DOCUMENT RESUME

ED 249 269 TM 840 622

AUTHOR	Hambleton, Ronald K.; Eignor, Daniel R.
TITLE	A Practitioner's Guide to Criterion-Referenced Test Development, Validation, and Test Score Usage (Second Edition). Laboratory of Psychometric and Evaluation Research Report No. 70.
INSTITUTION	Massachusetts Univ., Amherst. Laboratory of Psychometric and Evaluative Research.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
PUB DATE	10 Mar 79
NOTE	550p.
PUB TYPE	Guides - Classroom Use - Guides (For Teachers) (052) -- Tests/Evaluation Instruments (160)
EDRS PRICE	MF02/PC22 Plus Postage.
DESCRIPTORS	*Criterion Referenced Tests; *Cutting Scores; *Evaluation Methods; Mastery Tests; Models; Program Design; Research and Development; Scoring; *Test Construction; *Testing; Test Items; Test Norms; Test Reliability; *Test Results; Test Use; Test Validity
IDENTIFIERS	*Standard Setting

ABSTRACT

This instructional training package introduces practitioners to methods for developing, validating, using, and reporting criterion-referenced tests. It provides a comprehensive presentation of criterion-referenced testing technology. The package emphasizes the most recent substantive and technological advances in the field that are both important and relatively easy to use. The 10 units of instruction are: (1) "Introduction to Criterion-Referenced Testing"; (2) "Preparation of Objectives and Test Items"; (3) "Assessment of Content Validity"; (4) "Test Assembly and Administration"; (5) "Reliability, Validity and Norms"; (6) "Issues and Methods for Standard-Setting"; (7) "Criterion-Referenced Test and Test Manual Evaluations"; (8) "Use and Reporting of Test Score Information"; (9) "Design of Criterion-Referenced Testing Programs--Two Examples"; and (10) "New Developments and Areas for Further Research." Each unit is divided into sections: a unit overview; an introduction to covered topics; relevant technical materials and examples; occasional optional materials; and cited references. Some units have additional references for further study. Flow-charts, figures, and tables are included whenever possible. (Author/BS)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

ED249269

A Practitioner's Guide to Criterion-Referenced Test Development, Validation, and Test Score Usage[1,2]

(Second Edition)

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

M. M. Rogers

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM840 622

---

[1] The project reported herein was supported, in part, by a grant from the National Institute of Education, Department of Health, Education, and Welfare. However, the opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by the National Institute of Education should be inferred.

[2] Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1979. (2nd edition)

-March 10, 1979-

Introductory Comments

This instructional training package was prepared to introduce practitioners to methods for developing and validating criterion-referenced tests and to methods for using and reporting criterion-referenced test score information. In preparing the document we attempted to accomplish three goals:

1. Provide a comprehensive presentation of criterion-referenced testing technology;
2. Emphasize the most recent substantive and technological advances in the field;
3. Emphasize criterion-referenced testing contributions that are both important and relatively easy to use.

Material in the Practitioner's Guidebook is organized into ten units of instruction. The ten unit titles are:

1. Introduction to Criterion-Referenced Testing
2. Preparation of Objectives and Test Items
3. Assessment of Content Validity
4. Test Assembly and Administration
5. Reliability, Validity and Norms
6. Issues and Methods for Standard-Setting
7. Criterion-Referenced Test and Test Manual Evaluations
8. Use and Reporting of Test Score Information
9. Design of Criterion-Referenced Testing Programs—Two Examples
10. New Developments and Areas for Further Research

Each unit is divided into several sections. The first two usually provide an overview to the unit and an introduction to the topics which will be covered in the unit. The remaining sections provide, in a logical sequence, relevant technical material and examples. The selection of content reflects our bias toward criterion-referenced test methods which are fairly straightforward to understand, address satisfactorily the problems at hand, and are relatively easy to apply. In some instances it was necessary to violate this guideline when a method meeting our guideline was not available. Also, in a small number of instances, we included optional material which we felt may be of interest to practitioners. These sections are marked by an "*". They can be skipped without any loss in continuity.

The final section of each unit includes a list of cited references. In some units, we included a second list of references to facilitate further study of content introduced in the unit and/or special topics. We included flow-charts, figures, and tables whenever possible to improve the readability of our materials.

Many improvements have been made in the second edition of our Practitioner's Guidebook.

1. A slightly different model for developing criterion-referenced tests is proposed.
2. Many new examples of domain specifications prepared by curriculum specialists, teachers, and ourselves are included.
3. New methods and rating forms are offered for conducting content validation studies.
4. The material on approaches for assessing test score reliability is updated and tables offered to facilitate the determination of test length.

5. A unit on standard-setting is now available and the material in it is reflective of current issues and advancements in the area.

6. Guidelines and a rating form for evaluating criterion-referenced tests and test manuals are improved.

The present document is lengthy. Still, we are certain that several additional topics should be included in a third edition:

1. We limited our discussion to the construction and uses of paper and pencil tests. (In a later edition, we plan to add several units on performance testing.)

2. The present document includes no references to the measurement of affective outcomes. (The interested reader is referred to Popham's Criterion-Referenced Measurement, Prentice-Hall, 1978.)

3. We plan to include some case histories of teacher, school, district, and state-wide efforts to produce criterion-referenced tests.

4. A unit on descriptive and inferential statistics for test developers will be added. Also, we will include some material on Bayesian statistical methods and decision theory.

5. Methods for studying criterion-referenced test item and test score bias will be prepared as soon as sufficient information is available on the topic. For the moment, we recommend interested individuals apply the methods being advanced for use with norm-referenced tests.

We would like to hear from individuals who use the materials. Specifically, we would be interested in:

1. Your general comments about the materials;

2. Areas where you feel we have been incomplete, ambiguous, or inaccurate;

3. Your suggestions for expanding the material.

We expect that the Practitioner's Guidebook will be published in 1980. However, until the materials are published, the Northwest Regional Educational Laboratory (in Portland, Oregon) has agreed to distribute the document through its Clearinghouse.

Unit 1

Introduction to Criterion-Referenced Testing

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

Table of Contents

Page

1.0 Overview of the Unit. 1

1.1 Introduction. 2

1.2 Minimum Competency Tests. 8

1.3 Comparison of Norm-Referenced Tests and Criterion-Referenced Tests. 12

1.4 Shortcomings of Norm-Referenced Tests in Program Evaluation 15

1.5 Developing and Validating Criterion-Referenced Tests. 21

1.6 Using and Reporting Criterion-Referenced Test Score Information. 24

1.7 References. 26

1.0 Overview of the Unit

This unit was prepared to introduce readers to the topic of criterion-referenced test technology. Specifically, we will (1) provide some background, (2) consider the issue of definitions, (3) address the need for a theory and practice of criterion-referenced tests, (4) compare norm-referenced tests and criterion-referenced tests (5) consider shortcomings of norm-referenced tests, and (6) introduce a framework for studying the remainder of the units. With regard to point (6), we will introduce a twelve step model for developing and validating criterion-referenced tests. Also, we will introduce a framework for using and reporting criterion-referenced test score information.

1.1 Introduction

Glaser (1963) and Popham and Husek (1969) were the first to introduce and to popularize the field of criterion-referenced testing. Their motive was to provide the kind of test score information needed to make a variety of individual and programmatic decisions arising in objectives-based instructional programs. Norm-referenced tests were seen as less than ideal for providing the desired kind of test score information.

Presently, there are millions of students at all levels of education taking criterion-referenced tests. Criterion-referenced tests are used to monitor individual progress through objectives-based instructional programs, to diagnose learning deficiencies, to evaluate educational and social action programs, and to assess competencies on various certification and licensing examinations. There are many more uses as well.

Unfortunately, until recently (Hambleton and Eignor, 1978, 1979; Millman, 1974; Popham, 1978a), there have been few reliable guidelines for test construction, test assessment, and test score interpretation, and this in turn has hampered effective usage of criterion-referenced tests. Over the years, standard procedures for testing and measurement within a norm-referenced framework have become well-known to educators; however, these procedures are much less appropriate when the questions being asked concern what examinees can and cannot do (Glaser, 1963; Hambleton and Novick, 1973; Popham and Husek, 1969). Norm-referenced tests are constructed, principally, to facilitate the comparison of individuals (or groups) with one another or with respect to a norm group on the ability measured by a test.

Criterion-referenced tests are constructed to permit the interpretation of individual (and group) test scores relative to a set of objectives. Perloff, Perloff, and Sussna (1976) noted "the first recorded instance of evaluation occurred when man, woman and serpent were punished for having engaged in acts which apparently had not been among the objectives defined by the program circumscribing their existence." They might have added that the "assessment measure" was a criterion-referenced test. Adam's and Eve's behaviors relative to some stated objectives were compared to performance standards and found to be deficient. (Unfortunately, the combined failure of Adam and Eve on that single criterion-referenced test has had a long range effect on the rest of us. Fortunately, such long-lasting and far-reaching results of a person's criterion-referenced test scores are unusual.)

As an alternative to norm-referenced tests, criterion-referenced tests were introduced. Criterion-referenced tests are intended to meet the testing and measurement requirements in objectives-based instructional programs, competency-based certification programs, and numerous other situations where someone is interested in the performance of examinees relative to a set of objectives or competencies.

The last time anyone bothered to count, there were over 600 references on the topic of criterion-referenced testing. Unfortunately, there are almost as many ideas about what a criterion-referenced test is as there are contributors to the field. Ross Traub put it best when he suggested that some new contributions were akin to "stirring muddy water."

One of the major sources of confusion is over the word "criterion." For many individuals, it refers to a performance standard, a minimum proficiency level, or a cut-off score. But it is clear from the two most influential criterion-referenced testing papers in the 1960's (Glaser, 1963; Popham and Husek, 1969) that these writers used the word "criterion" to refer to a "domain of behaviors." These authors were interested in referencing examinee test performance to a well-defined domain of behaviors measuring an objective or competency. For further discussion of criterion-referenced test definitions, the interested reader is referred to Donlon (1974), Hambleton and Novick (1973), Millman (1974), and Popham (1978a).

Popham (1978a) provides the definition we prefer and the one we will work with in our instructional materials. It is:

> A criterion-referenced test is used to ascertain an individual's status [referred to as a domain score] with respect to a well-defined behavior domain.

It is often the case that a criterion-referenced test will measure more than a single objective. If so, items within the test are organized into non-overlapping subtests corresponding to the objectives measured by the test. Popham's definition is similar to the one offered by Millman (1974) and others for a domain-referenced test. The term, "domain-referenced test," is a good one because it is descriptive and therefore less apt to be misunderstood by practitioners. However, we agree with W. James Popham (one of the leading contributors to the field of criterion-referenced testing and a strong advocate for criterion-referenced testing) on the matter of which "test label" is the most useful. Presently, there is considerable public support for the term, "criterion-referenced tests" and therefore we feel it would be unfortunate if a new campaign had to be initiated by educators for the term, "domain-referenced tests." It is true though that few so-called "criterion-referenced tests" could satisfy the demands required by the criterion-referenced test definition offered above.

Currently, there is confusion over the differences among three kinds of tests -- criterion-referenced tests, domain-referenced tests, and objectives-referenced tests. If Popham's definition of a criterion-referenced test is adopted, there is no essential difference between criterion-referenced tests and domain-referenced tests. Objectives-referenced tests are tests consisting of items that are matched to objectives. The primary distinction between criterion-referenced tests and objectives-referenced tests is as follows: In a criterion-referenced test, the items are a representative set of items from a clearly defined domain of behaviors measuring an objective, whereas, with an objectives-referenced test, no domain of behaviors is specified, and items are not considered to be representative of any behavior domain. This distinction has important implications for the kinds of generalizations that can be made from criterion-referenced test scores as compared to objectives-referenced test scores and is the reason we prefer Popham's definition. It is interesting to note that most (if not all) commercially prepared "criterion-referenced tests" on the market today, would be called "objectives-referenced tests" if Popham's definition for a criterion-referenced test is adopted.

With the availability of a test theory for norm-referenced measurements, procedures exist for constructing appropriate measuring instruments, i.e., norm-referenced tests. Since the primary purpose for norm-referenced tests and criterion-referenced tests are fundamentally different, it is not surprising that a different theory and practice of testing is needed to handle the problem of testing to assess competence. It should be noted that a norm-referenced test can be used for criterion-referenced measurement, albeit with some difficulty, since the selection of items is such that many objectives will very likely not be covered on the test or, at best, will be covered with only a few items. It has been noted by at least two writers (Millman, 1974; Traub, 1972) that when items in a norm-referenced test can be matched to objectives, criterion-referenced interpretations of the scores are possible, although they are quite limited in generalizability.

A criterion-referenced test constructed by procedures especially designed to facilitate criterion-referenced measurement can be and sometimes is used to make norm-referenced measurements. However, a criterion-referenced test is not constructed specifically to maximize the variability of test scores (whereas a norm-referenced test is). Thus, since the distribution of scores on a criterion-referenced test will tend to be more homogeneous, it is obvious that such a test will be less useful for ordering individuals on the measured ability. In summary, a norm-referenced test can be used to make criterion-referenced measurements and a criterion-referenced test can be used to make norm-referenced measurements, but neither usage will be particularly satisfactory (Hambleton and Novick, 1973).

It has been argued by some test developers that to refer to tests as either norm-referenced or criterion-referenced tests may be misleading since measurements obtained from either testing instrument can be given a norm-referenced interpretation, criterion-referenced interpretation, or both. The important distinction made was that between norm-referenced measurement and criterion-referenced measurement (Glaser, 1963; Hambleton and Novick, 1973). From a historical perspective, this distinction was important since a methodology for constructing criterion-referenced tests did *not* exist, at least at the time of Glaser's article. Criterion-referenced tests were constructed in the same manner as norm-referenced tests, and as pointed out above, the usage was *not* satisfactory. However, in view of recent developments in the field, it would be correct to refer to a test as either *criterion-referenced* or *norm-referenced*. In fact, given the operational definitions, the distinctions between criterion-referenced tests and norm-referenced tests are both unambiguous and meaningful.

Of course, not all educators agree on the usefulness of criterion-referenced tests (see, for example, the debates between Block [1971] and Ebel [1971] and between Popham [1978b] and Ebel [1978]). Our position is that criterion-referenced tests can serve a wide variety of uses, and their usefulness will be enhanced through knowledge and understanding of technical developments which address their proper construction, validation, and usage.

BEST COPY AVAILABLE

1.2 Minimum Competency Tests

The establishment of minimum competency testing programs in elementary and secondary schools, and for many professions, has reached immense proportions (or epidemic proportions, if you view the trend negatively). For example, well over half (33 to be exact) of the states have passed legislation requiring assessment of the "competence" of their elementary and high school students (Pipho, 1978). Further, many of these states require that students demonstrate at least a minimum level of performance on a set of competencies in order to receive a high school graduation diploma. Why are so many state legislatures mandating minimum competency testing? It appears that it is to discourage schools from the practice of promoting all students and awarding high school graduation diplomas based on school attendance only. It is common for legislators and parents to say that minimum requirements in the "basic skills" must be set for students to graduate with a diploma which has some meaning.

The rapidity of change in school, district, and statewide testing programs and the demand for high quality tests has dictated that substantial research and development work be undertaken. Included among the more important research and development topics are: Identification and definition of competencies, management of competency testing programs, development and validation of competency tests, methods of determining standards, and uses and interpretations of competency test scores.

Competency testing technology would be in an embryonic stage were it not for the work done in developing a criterion-referenced testing technology since the late 1960's. A competency test is simply a particular kind of criterion-referenced test and therefore, like a criterion-referenced test, it must be developed and used in ways somewhat different to better-known norm-referenced tests. All (or nearly all) which follows in this Practitioner's Guidebook will apply equally well to both criterion-referenced tests, and what have come to be known as competency and minimum competency tests.

Perhaps we should begin with a definition of a competency test:

> A competency test is designed to determine an examinee's level of performance relative to each competency being measured. Each competency is described by a well-defined behavior domain (Hambleton and Eignor, 1979).

Clearly, competency tests and criterion-referenced tests are equivalent. Perhaps the only differences are the contexts in which the terms are used and the characteristics measured by the tests. However, there is no need for two expressions. The definition makes clear that the purpose of a competency test is to provide information about an individual examinee's level of performance on each competency which is measured by a test. There will be as many test scores as there are competencies measured by a test. Also, competencies are clearly written so that there will be a high level of agreement among users of the test about the content (behaviors) defining the competency. This desirable goal can be accomplished through the use of "domain specifications" (Popham, 1978a). This term will be described in more detail later. There is one other point. There is nothing inherent in the definition of a competency test which requires test scores to be compared to "standards."

In fact, the percentage scores (reported by competency) provide excellent descriptive information about examinee performance. Since it is common, however, to interpret examinee test performance relative to standards (an examinee who scores equal to or above a standard set at 70% [say] on the set of test items included in a competency test is described as a "master" or "competent"), it is necessary to introduce a new term, "minimum competency testing."

> A minimum competency test is designed to determine whether an examinee has reached a prespecified level of performance relative to each competency being measured. The "prespecified level" or "standard" may vary from one competency to the next. Also, each competency is described by a well-defined behavior domain.

A "standard" (sometimes it is called a "cut-off score" or a "minimum proficiency level") is a point on a test score scale which is used to separate examinees into two categories, each reflecting a different level of proficiency relative to the competency measured by the test under consideration. It is common to assign labels such as "master" or "competent" to those persons in the higher-scoring category and "non-master" or "competent" to those persons in the lower-scoring category. Note that if a test measures more than a single competency and if examinees are to be classified into competency categories based on their performance on each set of items measuring a competency, as is often the case, a standard is set for each competency measured by the test. There can be as many competency decisions as there are competencies measured by the test.

From the definitions above, it is clear that minimum competency tests are a special type of competency test (tests where standards are introduced to interpret examinee performance) and as we mentioned earlier, competency tests are a special type of criterion-referenced test (i.e., those tests which are used usually in certification and licensing situations).

1.3 Comparison of Norm-Referenced Tests and Criterion-Referenced Tests

Similarities and differences between norm-referenced tests and criterion-referenced tests are summarized below under nine topics.

Purpose

A norm-referenced test is designed to facilitate comparisons among examinees on the ability being measured.

A criterion-referenced test is designed to assess an examinee's level of performance relative to a well-defined behavior domain.

Test Development Method

For a norm-referenced test, a test blueprint is prepared and items are written according to the blueprint. An important factor in item selection is the statistical properties of the test items (item difficulty and discrimination). In general, items of moderate difficulty (p-values in the range .30 to .70) and high discriminating power (point biserial correlations over .30) are the most likely to be selected for inclusion in a test because they contribute substantially to test score variance. Test reliability and validity will, generally, be higher when test score variance is high.

For a criterion-referenced test, domain specifications are prepared and items written to measure the domain specifications. Test items are selected for a criterion-referenced test if they are "reflective" of the domain they were written to measure and if they can serve as a "representative" set of test items defined by the domain specification. (A domain specification represents an attempt to clearly define the behavior domain associated with a particular objective or competency.)

Measurement Scales

The norm-referenced test score scale is anchored in the middle (the average level of group performance).

For criterion-referenced test score scales, the anchor points are two in number and located at the ends of the scale (0% and 100%).

Test Score Uses

Norm-referenced test scores are often used to make comparisons among examinees or to handle "fixed quota" selection problems (i.e., the problem when there is a fixed number of "vacancies" and the number of applicants exceeds the number of vacancies).

Criterion-referenced test scores are used (1) to make descriptive statements about what examinees can do, (2) to make instructional decisions, and (3) to evaluate programs and their effectiveness. Examinees are judged primarily on their own merits. There are instances where examinees (or groups) may be compared with one another, but this is not a primary use of the scores. Criterion-referenced tests are often used in "quota-free" selection problems (i.e., situations when no limits are placed on the number of examinees receiving a "passing score").

An important point to note is that both norm-referenced tests and criterion-referenced tests "sort" individuals. However, norm-referenced tests are used to sort examinees according to their performance on the test and criterion-referenced tests are used to sort examinees into groups according to their mastery or non-mastery of skills measured by a test.

Test Score Generalizability

There is seldom interest in making generalizations from norm-referenced test scores. Usually, the job is completed when test scores are compared to appropriate norms tables.

With criterion-referenced test scores, the matter of generalizability is important. Seldom would anyone be content to interpret an examinee's score in terms of the specific items on a test. (Incidentally, this is all that can be appropriately done with scores obtained from objectives-referenced tests.) If the objective measured by a test is clear, and if items are selected to be representative of the behavior domain defining the objective, examinee test performance on a set of items included in the test can be generalized to test performance in the larger domain of behaviors. Strong criterion-referenced test score interpretations of the kind just described are usually of interest to criterion-referenced test users. (So much so that they usually make a strong interpretation whether justified or not!)

Specificity of Test Score Information

A norm-referenced test provides a summary of a somewhat abstract area of achievement.

A criterion-referenced test provides very specific and detailed information about a clearly defined area of achievement.

Instructional Applications

Users of norm-referenced tests espouse the view that learning is a complex process consisting of concepts and relationships organized in a hierarchical arrangement.

Users of criterion-referenced tests are endorsing the notion that things learned can be separated into discrete categories (referred to as objectives, skills, or competencies).

Reliability and Validity Issues

For both types of tests, reliability and validity considerations are important. However, since test score reliability and validity are also specific to the intended uses of the scores, and since norm-referenced test scores and criterion-referenced test scores are used to address different types of problems, it is not surprising that approaches for assessing reliability and validity will differ with each type of test.

Norms

Norms tables are of central importance with norm-referenced tests. Norms can also be of value when interpreting individual and group criterion-referenced test scores.

1.4 Shortcomings of Norm-Referenced Tests in Program Evaluation[1]

Evaluators are confronted with questions such as, "What is the average level of performance on a particular set of mathematics objectives for a specified group of individuals?" or "How much has a particular group learned from a special reading program?" A common question asked of Title I program evaluators is "Have X% of a group of participating students "mastered" over Y% of the reading program objectives?" It is common practice for individuals presented with these questions to turn to one of Oscar Buros' Mental Measurements Yearbooks (the eighth edition was published in 1978) and search for a suitable assessment instrument. But the search is likely to be a long and frustrating one. The great majority (probably over 95%) of the instruments found in the Yearbooks are norm-referenced instruments. That is, the instruments were designed primarily to permit comparisons of one individual with another on the construct or ability measured by the instrument. Grade-equivalent scores, age-equivalent scores, percentile ranks, and standard scores are common ways of reporting individual test performance. All reporting methods permit comparisons among individuals, but they provide little or no information relative to important questions such as, "What can an individual (or group) do?"

Because they are used so frequently, you might think (or be tempted to conclude) that norm-referenced tests provide excellent indicators of program effectiveness. Federal agencies, school boards, program evaluators, and even parent groups often request that they be administered. Why? Well, certainly it could be argued that they are objective, and usually norm-referenced tests are developed and distributed by companies with years of experience in the area of testing. Also, the costs are usually low (at least when compared to the costs of a program developing and validating its own evaluative instrument). But, and this is the critical point, it is so difficult to purchase a norm-referenced test where the content of the test will closely approximate the objectives or goals of some specific program.

---

[1] From a paper by Hambleton, R.K., & Gifford, J. C. Development and use of criterion-referenced tests to evaluate program effectiveness. Laboratory of Psychometric and Evaluative Research Report No. 52. Amherst, MA: School of Education, University of Massachusetts, 1977.

Let us back up now and look more closely at the construction and uses of norm-referenced tests. A norm-referenced test is a test that is designed to facilitate the comparison of individuals with respect to one another, or some appropriately chosen norm group. Since the major purpose of norm-referenced tests is to facilitate comparisons, the information obtained from a norm-referenced test is information concerning an individual and is best used to make decisions about individuals. Consequently, norm-referenced tests are often used for selecting, placing, and counseling individuals. In order for a norm-referenced test to be effective in making meaningful comparisons, it is important that there be variability in individual responses. This is accomplished by selecting test items from a larger pool of available test items for an instrument because they have moderate item difficulty levels (typically in the range .30 to .70) and moderate to high discrimination indices (point biserial correlations over about .30). Other things being equal, items so selected will tend to maximize test score variability. Variability spreads the examinees over the ability scale and allows the user to make meaningful comparative statements about an individual in terms of the group as a whole. Without variability, all examinees would be receiving the same scores and no useful information would be obtained. In norm-referenced testing, all meaning is dependent entirely on the comparison of the individual to others.

The most commonly used norm-referenced tests are those prepared by publishers to be used extensively throughout the country. Test items are chosen to measure content or goals common to a wide sample and types of programs. The norm-referenced test provides an overview of an individual's relative ability in a rather broad content area. The scores are reported as raw scores and one or more derived scores (for example, percentiles, age- or grade-equivalent scores, and standard scores). The raw scores alone have very little meaning. Inferences cannot be made as to what the student knows or does not know. The derived scores give specific information concerning the relation of an individual's knowledge or ability to that of a particular reference group.

With the increase of concern for accountability and the efficient use of tax dollars, program evaluation has taken a position of tremendous importance. Government agencies, for example, need the kind of information that will enable them to make the most effective decisions. One of the most commonly used measures of program effectiveness is the student test score, in particular, the norm-referenced standardized test score. If this is to be the case, it is crucial that the tools chosen for evaluation be ideally suited for answering the questions an evaluator asks. For several reasons norm-referenced tests are not well-suited for the measurement of program effectiveness. However, there are others who take a different view (Ebel, 1971, 1978).

One shortcoming of norm-referenced tests in program evaluation is the discrepancy between the content covered by a test and the content of a program that is being evaluated. Reasons for the discrepancy relate to the basic construction and the misuse of norm-referenced tests. The tests that are most commonly used in evaluations are used nationwide and are based on an amalgamation of objectives of programs from all over the country. Each program has different objectives and different times when the instruction of particular objectives occurs. The overlap of program objectives and test objectives will not be complete and the degree of overlap will change from program to program. This is particularly true in compensatory educational programs, where the objectives may be more basic and specific than the general objectives reflected in norm-referenced tests.

It is often hard to find a standardized achievement measure where the content covered by the measure closely matches the content goals of a particular program being evaluated. Therefore, any evidence from the achievement measure can always be discarded. When the match between test content and program content is low, we have nothing of value. Since each program curriculum typically reflects the people teaching the program and their priorities and emphases, this "mismatch" is commonly encountered in program evaluation studies.

A second cause of the discrepancy between test content and program objectives arises directly from a major purpose of norm-referenced tests, i.e., to compare an individual's performance, knowledge, or skill to that of some reference group. In order to effectively obtain this type of information from a test, the test must be constructed with that purpose in mind. Consequently, norm-referenced tests consist of test items that contribute most to maximizing test score variability. In the process of choosing items that contribute variability, those contributing low variability are eliminated. It is clear that items tapping concepts taught successfully by a great number of teachers will contribute little to test score variability (most students will answer the items correctly) and will be eliminated, while the items measuring pure reasoning ability will have greater variability and will be retained. As a result of the process, the test begins to look less like an achievement test and more like an aptitude test. The process of item selection puts a distance between the curriculum of the educational program and the tool used to evaluate it. The test would be sensitive to the aptitude of the individuals rather than the effectiveness of the instruction. If an instrument is to be sensitive to the learning process, its content must be very carefully matched to that of the program. It is being said more and more that norm-referenced tests function like IQ tests. Test items where performance is high (perhaps reflecting areas of successful teaching) are typically removed because they fail to discriminate. In other words, many school-related skills are systematically eliminated. What we are left with is variation due to the effects of "non-school-related variables."

Presently, many of the programs to be evaluated are innovative. Not only are the instructional methods different, but often the goals and objectives of the program are different from those of the traditional program. As a result, a score doesn't represent knowledge in terms of the instruction. It is a mistake to judge an innovative program according to the standards of a traditional program. The effectiveness of a program cannot be measured by a tool that has been developed to measure something else.

Other problems with norm-referenced tests result not from the basic construction but from the use of the tests. In many cases, the program to be evaluated deals with a population that is not reflected by the norm group of the test. This has implications for the interpretation of scores for many types of compensatory and special educational programs.

There is yet another problem. Standardized or published norm-referenced tests can be criticized also on the grounds that they are too general. Of course, they have this feature to give them broad appeal, but the more general the test, the easier it is for people to see what they want in the results.

1.5 Developing and Validating Criterion-Referenced Tests

Figure 1.5.1 from Hambleton and Eignor (1979) presents a twelve step model for developing and validating criterion-referenced tests. The importance of each step in the model depends upon the size and scope of the test development and validation project. An agency with the responsibility of producing a criterion-referenced test for state-wide use will proceed through the steps in a rather different fashion than will a small consulting firm or a teacher producing a classroom test on a very limited budget.

In brief, the twelve steps are as follows:

Step 1--Objectives must be prepared or selected before the test development process can begin.

Step 2--Test specifications are needed to clarify the test's purposes, desirable item formats, number of test items, instructions to item writers, etc.

Step 3--Items are prepared to measure objectives included in the test (or tests, if there are going to be parallel-forms, or levels of a test varying in difficulty).

Step 4--Initial editing of items is completed by the individuals writing them.

Step 5--A systematic assessment of items prepared in steps 2 and 3 is conducted to determine the item validities. Essentially, the task is to determine the content validity of the test items.

Step 6--Based on the data from step 5, it is possible to do further item editing, and in some instances, discard items that do not adequately measure the objectives they were written to measure.

Step 7--The test (or tests) must be assembled.

Step 8--A method for setting standards to interpret examinee test performance is selected, and implemented.

Step 9--The test (or tests) can be administered.

Step 10-Data addressing reliability, validity, and norms should be collected and analyzed.

Step 11--A user's manual and a technical manual should be prepared.

Step 12--This step is included to reinforce the point that it is necessary, in an on-going way, to compile technical data on the test items and tests as they are used in different situations with different examinee populations.

The next six units (units two to seven) will provide details for successfully completing each of the steps in the test development and validation model presented in Figure 1.5.1. The chart below summarizes the location of instructional material on each of the twelve steps in the units of instruction which follow:

Steps	Unit
1,2,3,4,	2
5,6	3
7,9	4
8	6
10	5
11,12	7

1. Writing and/or Selection of Objectives
2. Preparation of Test Specifications (for example, Available Time, Selection of Objectives to be Measured by a Test, Number of Test Items/Objective, Appropriate Vocabulary, Method of Scoring)
3. Writing Test Items "Matched" to Objectives
4. Preliminary Review of Test Items
5. Determination of Content Validity of the Test Items
    - (a) Involvement of Content Specialists
    - (b) Collection and Analysis of Examinee Item Response Data
6. Additional Editing of Test Items
7. Test Assembly
    - (a) Determination of Number of Test Items/Objective
    - (b) Test Item Selection
    - (c) Preparation of Directions and Sample Questions
    - (d) Layout and Test Booklet Preparation
    - (e) Preparation of Scoring Keys
    - (f) Preparation of Answer Sheets
8. Standard Setting for Interpreting Examinee Test Performance
9. Test Administration
10. Assessment of Test Score Reliability and Validity; Compilation of Test Score Norms (Optional)
11. Preparation of a User's Manual and a Technical Manual
12. Periodic Collection of Additional Technical Information

Figure 1.5.1. Steps for Developing and Validating Criterion-Referenced Tests.

1.6 Using and Reporting Criterion-Referenced Test Score Information

Figure 1.6.1 outlines the content of Unit 8. Specifically, we will consider two primary uses of criterion-referenced test scores: (1) Domain-score estimation, and (2) mastery status determination. We also discuss, but to a lesser extent, the use of criterion-referenced test scores for program evaluation. For each use, we will consider appropriate methods for applying criterion-referenced tests. Finally, we consider a number of ways of reporting test score information, and we consider the use of criterion-referenced tests for grading purposes.

1. Uses of Criterion-Referenced Tests
2. Domain-Score Estimation
   (a) Selection of an Estimation Method
3. Mastery Status Determination
   (a) Selection of a Decision Model
   (b) Loss Specification
4. Reporting of the Information
   (a) Individual Level
   (b) Group Level
   (c) Program Evaluation
5. Criterion-Referenced Grading

Figure 1.6.1 Using and Reporting Criterion-Referenced Test Score Information

1.7 References

Block, J. H. Criterion-referenced measurements: Potential. School Review, 1971, 69, 289-298.

Donlon, T. F. Some needs for clearer terminology in criterion-referenced testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

Ebel, R. L. Criterion-referenced measurements: Limitations. School Review, 1971, 69, 282-288.

Ebel, R. L. The case for norm-referenced measurements. Educational Researcher, 1978, 15, 321-327.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Hambleton, R. K., & Eignor, D. R. Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard-setting. In R. Jaeger & C. Tittle (Eds.), Minimum competency testing. (Approx. Title) Berkeley, California: McCutchan Publishing Co., 1979.

Glaser, R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-521.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C. W., Alkin, M. C., & Popham, W. J. Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Co., 1974.

Perloff, R., Perloff, E., & Sussna, E. Program evaluation. Annual Review of Psychology, 1976, 27, 569-594.

Pipho, C. Minimum competency testing in 1978: A look at state standards. Phi Delta Kappan, 1978, 59, No. 9 (May), 585-587.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978. (a)

Popham, W. J. The case for criterion-referenced measurements. Educational Researcher, 1978, 7, 6-10. (b)

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Traub, R. E. Criterion-referenced measurement: Something old and something new. A paper prepared for an invited public address at the University of Victoria, 1972.

Unit 2

Preparation of Objectives and Test Items

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

Table of Contents

Page

2.0 Overview of the Unit 1

2.1 Introduction 2

2.2 Development of Objectives 5

2.3 Domain Specifications 10

2.4 Examples of Domain Specifications 22

2.5 Item Forms Analysis 56

2.6 Examples of Item Forms Analysis 61

2.7 Flowchart of the Process of Developing Objectives 68

2.8 Objective Banks 69

2.9 Preparation of Test Specifications 70

2.10 Preparation of Test Items 77

2.11 Editing Test Items 96

2.12 References 100

2.12.1 References Cited 100
2.12.2 References for Further Study 102
2.12.3 Measurement and Evaluation Textbooks 103

2.0 Overview of the Unit

This unit covers the first four steps of the Criterion-Referenced Test Development and Validation Model presented in Unit 1. These steps are:

1. Writing and/or Selection of Objectives
2. Preparation of Test Specifications
3. Writing Test Items "Matched" to Objectives
4. Preliminary Review of Test Items

2.1. Introduction

In Unit 2, we will discuss both research and procedures directed toward the preparation of objectives and test items for criterion-referenced tests. Before offering a relevant set of procedures for the development of objectives and items, it is necessary to introduce and discuss some important background information that will help the reader better understand the interrelationships among the procedures to be discussed.

Popham's (1978a) definition of a criterion-referenced test is an excellent starting point for this discussion. It is as follows:

> A criterion-referenced test is used to ascertain an individual's status [referred to as a domain score] with respect to a well-defined behavior domain.

Once the domain of relevant behaviors has been defined, test items are written to measure behaviors in the domain, and a test is formed. From the test results, a test practitioner usually desires to make an inference about an examinee's level of performance relative to the domain of behaviors. A valid inference can be made if two conditions[1] are met:

1. The behavior domain is clearly and completely specified.
2. A random sample (or stratified random sample) of tasks from the domain is measured by the test.

In attempting to achieve these two conditions, several problems arise. From a technical standpoint, for a random sample to be taken (step 2), it is not enough for the domain to be well-defined, it must also be completely defined. For certain subject domains, this complete definition of domain may be impossible, and a compromise must be reached. Weaker inferential procedures must be considered.

---

[1] Actually, if the test items have been calibrated using one of the many latent trait models, the second condition need not be satisfied (Hambleton, 1979).

This issue of domain specification was described by Traub (1975) as the concern for domain sampling validity. Domain sampling validity concerns the adequacy of the tasks contained in a test as a sample of the tasks in the whole domain. According to Traub:

> It is the kind of validity that establishes the basis for one kind of inference from observed level of performance on a test to probable level of performance on all tasks contained in the domain.

Traub further distinguishes two varieties of domain sampling validity, and these two varieties are directly related to the procedures discussed in this unit. Strong domain sampling validity can occur when the domain is explicitly and completely defined; weak domain sampling validity occurs when the domain of tasks cannot be defined explicitly, and so, must be defined implicitly. A domain of tasks is explicit if all the tasks in the domain are known; the description is implicit if not all the tasks are known, but a clear enough description of the domain exists to see how the tasks should arise. The most frequently used approach to forming explicit descriptions is to employ the use of item generation forms (the procedure is called item forms analysis), which is discussed in section 2.6.

A frequently used approach based upon an implicit description of the domain is through the use of domain specifications, a topic which will be discussed in section 2.3.

Strong domain sampling validity will be difficult to achieve, except in highly structured content areas such as mathematics.

Usually behavior domains cannot be explicitly defined. How can weak domain sampling validity be assessed in these situations? Cronbach (1971) and Cronbach *et al*. (1972) require that if a domain cannot be explicitly defined, that the implicit definition be clear enough so that ". . .qualified judges can agree as to whether any particular test item is included in the definition or ruled out by it." To be able to generalize from the sample of test items to the domain when the domain is implicitly defined requires a *replication* of the item writing task by different content specialists and a comparison of examinee scores on the two forms of the test constructed by the two groups of content specialists working independently with the same set of test specifications. The use of content specialists in establishing the validity of test items will be discussed in Unit 3.

In summary, the test constructor wants the test to be a random sample of the behaviors included in a domain so that he/she can make an inference about examinee performance on the domain, based on the test results. When the domain can be explicitly defined, such as through the use of item generation forms, *random sampling* can occur, and the desired *strong* inference can be made. When the domain is implicitly defined, such as through the use of domain specification procedures, *representative sampling* can occur, but a somewhat more limited inference can be made. Replication can, however, improve the strength of the inference.

2.2 Development of Objectives

The purpose of this section is not to debate the merits of using behavioral objectives. This has been a hot topic of debate for years (Popham, 1968; MacDonald & Wolfson, 1970; Allendoerfer, 1971; Forbes, 1971; Gagné, 1972; Ebel, 1973; Duchastel & Merrill, 1973; Kneller, 1972). For our purposes here, little could be accomplished by a review of the literature. We will offer three reasons why we feel behavioral objectives (or some variants) should be used, and then continue to the specifics that are pertinent to criterion-referenced test development. Behavioral objectives: (1) Serve as a mechanism for organizing a curriculum, (2) provide information to students about what is expected of them, and (3) provide a basis upon which to assess student performance. It is the third reason that is critical in the development of criterion-referenced tests. Behavioral objectives are a necessity as a starting point for setting up a criterion-referenced testing program. However, as the discussion that follows will show, behavioral objectives are not sufficient; more specification is needed.

Hopefully, the following historical development will provide a context in which to view the present state of usage of objectives in the development of criterion-referenced tests.[1] For the past few years, it has been a popular procedure for criterion-referenced test developers to write their objectives in "behavioral terms." However, while behavioral objectives have some desirable features (for example, they are relatively easy to produce), they often lack sufficient clarity to permit a clear determination of the domain of test items measuring the behaviors intended to be defined by an objective. For example, for even a simple mathematics objective such an adding two single-digit numbers, test developers would need information about the use of vertical versus horizontal format, the use of negative numbers, the number of test items to be placed on each page, etc. If the proper domain of test items measuring an objective is not clear, it is impossible to select a representative sample of test items from that domain. Since it is desired to interpret an examinee's test performance on the sample of test items measuring a particular objective as an estimate of that examinee's level of mastery in the larger domain of items measuring an objective, it is essential to have the domain of test items specified clearly, and to choose a representative sample of test items. When the domain of behaviors is not clearly spelled out, we have what Popham (1974) calls a "cloud-referenced test."

---

[1] Popham (1978b) offers an excellent summary of the development and specification of objectives.

A recent advancement that has proven more useful than the behavioral objective is the "amplified objective" (Popham, 1975). According to Millman (1974), "An amplified objective is an expanded statement of an educational goal which provides boundary specifications regarding testing situations, response alternatives and criterion of correctness." The importance of these additional guidelines added to a behavioral objective is that they help to define the relevant domain of test items. There is still some ambiguity in the domain definitions but the situation is considerably improved over using behavioral objectives. A sample behavioral objective and amplified objective are presented in Figure 2.2.1.

To further alleviate the ambiguity mentioned above in the use of amplified objectives, Popham (1975) presented a procedure for the development of domain specifications. Domain specifications can be viewed as a logical extension of amplified objectives, where there is also careful clarification of the content specified by an objective besides a

Figure 2.2.1. An Illustrative IOX Amplified Objective for a Third-Grade Level Reading Comprehension Skill[1]

DETERMINING SEQUENCE FROM TENSE AND WORDS THAT SIGNAL ORDER

Objective: The student will correctly identify the sequence of three sentences by determining order from tense and words that signify order.

Sample Item:

Directions. Read the three sentences. Then mark an "X" next to the answer that arranges the sentences in the proper order.

Example: A. Once there were only candles for lighting the home.

B. Later, there were dim electric lights.

C. Tesla thought of a way to make the electric lights brighter.

____a) A,C,B ____b) A,B,C ____c) C,A,B

Amplified Objective:

Testing Situation.

1. The student will be given three sentences and will identify their proper sequence on the basis of verb tenses and signal words.
2. Three sentences containing signal words, and/or changes in verb tense will be provided.
3. Vocabulary will be familiar to the third grader.

Response Alternatives.

1. Three possible orderings of the sentence will be given.
2. At least one distractor should not consist of a random ordering. It should maintain the first event as first, varying only the second and third events.
3. The other distractor may be any other incorrect ordering of the events.

Criterion of Correctness. The correct answer will be the order which can be determined on the basis of one of the following:

1. words that signify sequence, e.g., afterwards, finally, then, before, during, now, next, lastly, later, earlier, meanwhile, long ago, once;
2. verb tense (future, past, present).

---

[1] Reproduced from Popham (1978b, permission pending).

specification of boundary conditions, response alternatives and criteria for correctness. In section 2.3, the development of domain specifications will be discussed in greater detail.

Besides the use of the procedures for developing domain specifications being used with amplified objectives, similar concerns may be addressed with item forms analysis, or the development of item generation forms (Hively et al., 1973). In this situation, two requirements need to be met:

1. All the items which could be written from the content domain to be tested must be written (or known) in advance of the final selection process.
2. A random or stratified random sampling procedure must be used in the item selection process.

Item forms analysis assures that the above two requirements are met. An item form is actually a process having the following characteristics:

1. It generates items with a fixed syntactical structure.
2. It contains one or more variable elements.
3. It defines a class of item sentences by specifying the replacement sets for the variable elements.

Before discussing domain specifications and item forms in greater detail, two comments can be made. One, the reader should link the above discussion of domain specifications and item forms back to the introduction, and the distinction set up between strong and weak domain sampling validity. Two, as pointed out by Hively et al (1973), the use of item forms analysis is relevant only for highly structured subject areas such as mathematics. Popham's domain specification procedure appears to be relevant for a much wider variety of subject areas, and hence, in the sections to follow, more emphasis will be placed on domain specifications.

It should be noted that there appear to be several other promising ways for defining a "behavior domain." Besides "domain specifications" and "item forms analysis," facet theory (Berk, 1978), item transformations (Anderson, 1972; Bormuth, 1970) and algorithms (Scandura, 1977) have been suggested. However, these last three methods will not be considered further here. The cited references above will guide the interested reader to further study of these important new developments.

2.3 Domain Specifications

Popham (1975, 1978a) has prepared a series of steps that allow the test developer to produce a domain specification. According to Popham (1978a):

> The most important attribute of a criterion-referenced test is that it provides a clear description of the class of behavior that the examinee can or cannot perform. In fact, this description of measured behavior constitutes the "criterion" to which the test is "referenced."

The steps Popham has developed help in describing the class of behaviors discussed above in the quote. According to Popham, a domain specification should have two desirable qualities when prepared:

1. The specification should be brief enough to be used by the developer, and at the same time,

2. the specification serving as the domain description should be stated so that it "sufficiently circumscribes the class of behaviors under consideration so that independent judges will register high agreement regarding whether particular test items do, in fact, measure the behavior described in that domain."

The following series of steps are those suggested by Popham (1975, 1978a). We have included together suggestions he has made in both of his books to arrive at an "all-encompassing" set of steps one might follow to produce a domain specification. Before presenting an outline of these steps and a further discussion of the relevance of each, two comments should be made. One, the domain specifications developed through utilization of Popham's procedure lead to a situation that Traub has described as having weak domain sampling validity. Only when the stimulus attributes (step 3c in Popham's procedure) can be described in totality can strong domain sampling validity be obtained. Second, in what follows, steps one and two are concerned with general considerations one must attend to before the actual domain specification can be prepared, which are described in step 3. With this in mind, an outline of Popham's (1975, 1978a) steps for the preparation of a domain specification is as follows.

1. Zeroing in on the behavior to be measured: Degree of generality
    a. instructional duration
    b. limited priorities
    c. item homogeneity
2. Selecting from competing domain alternatives
    a. transferability within domain alternatives
    b. transferability outside the domain
3. Component steps in actual domain specification preparation
    a. general description
    b. sample item
    c. stimulus attributes
    d. response attributes
    e. specification supplement

1. Zeroing in on the behavior to be measured: Degree of generality

The question that must be first answered in the ultimate quest for a clear domain specification is, according to Popham (1975), "How large a chunk of an individual's behavior should we set out to circumscribe?" There is a trade-off here between choosing a large area of behavior, thereby forcing minute detail in order to adequately specify the domain, and smaller areas of behavior which collectively may bring about more domain descriptions than people will ever use. Popham (1978a) argues for choosing smaller segments of examinees behavior because the more finite behaviors are easier to isolate and circumscribe. He offers three possible ways to aid in the choice of degree of generality: Instructional duration, limited priorities, and item homogeneity. These are offered as guidelines for choice and can be used separately or collectively. In reference to instructional duration, Popham (1975) states:

> One way of thinking about a domain is to consider the amount of instructional time it would typically take to get learners to display the behavior depicted in the domain description.

In other words, the size of the domain to be circumscribed can be dictated by how long it takes to instruct students in the domain. In reference to limiting priorities, (the second guideline) one can get a "handle" on domain magnitude by setting a limit on the number of domains to be used, and then making sure that the most important behaviors to be sought are incorporated in this group of limited domains. Finally, the domain generality issue can be resolved by setting as a limit the domain description that would yield one variety of homogeneous item. That is, the domain to be circumscribed can be only large enough that the items generated from the domain perform the same function. Sameness of function can be observed by looking at the similarity between items in content and format.

Suppose that by using one of the suggested guidelines, or some other practical method, the test developer has come up with that he/she feels is a suitably sized segment of behavior to be measured, a segment that can be adequately "circumscribed." In other words, the test developer has come to closure on the degree of generality of the behavior he/she is going to try to measure. The next problem is deciding on which of a number of measurement approaches to use to try to measure the domain.

2. Selecting from competing domain measurement alternatives

The problem to be addressed at this point is to decide upon one from a number of measurement approaches possible to assess the behavior under consideration. For example, if one is trying to assess a student's ability to add, there are numerous measurement possibilities: Numbers listed vertically ($\begin{array}{r}4\\ +\underline{2}\end{array}$), numbers listed horizontally (4 + 2), numbers in equation form (4 + 2 = x), or possibly in verbal form (Joe has 4 oranges, Anne gives him 2 more, how many does he have?). There are more possibilities for this very simple task; for complex tasks, the possibilities that can be written down are even less exhaustive. Popham (1975) warns that while it is enticing to combine all (or many) of the possibilities in a single domain, this would cause confusion as to what constituted the domain in the first place. One measurement procedure must be chosen and Popham offers that the measurement alternative chosen should be the most generalizable of the possibilities considered and also be the one most able to be transferred outside the particular domain to others. In reference to generalizability across alternatives, another way of looking at this would be to select the alternative that, when mastered, would be most likely to reflect mastery of the other possibilities. In reference to degree of transfer, the alternative that transfers the most to other skills, courses, etc., should be chosen. Once again, these are only potential guidelines to be used, there are no hard-and-fast procedures for selecting the measurement alternatives.

In sum, by selecting from the myriad of possible ways of measuring behavior the one form of measured behavior that is the most generalizable, we will have (Popham, 1975):

> ...the best of two worlds, that is, an adequate reflection of the attribute being assessed plus an understandable set of test results.

Thus far, the test developer has decided upon the degree of generality of the domain he/she wants to deal with, and has further chosen a particular assessment approach from a larger group. The next step is to actually generate the domain specifications.

3. Component steps in domain specification preparations

Popham (1975) makes the point here that differing approaches to the describing of a relevant domain vary in degree of detail. He offers a set of procedures that have been modified through use with the Instructional Objectives Exchange. Popham recognizes that they are far less "elaborate" than the procedures developed by Hively, et al. (1973) for describing a domain of behavior. The utility of Popham's approach, we feel, is generated out of being able to use his set of steps over a wide variety of subject domains, especially those that are less structured (i.e., humanities).

The first component of a domain specification to be prepared is a general description of exactly what the test purports to measure. This provides an overview of the behavior (or set of behaviors) that are described in detail later. While this component could be suitably called an objective, Popham prefers to call it a general description. In his 1978 book, he offers the following example of a general description for a CRT dealing with the scientific method.

> When given brief, previously unseen fictitious accounts of the research activities of natural and physical scientists, students will answer questions (keyed to the accounts) calling for the identification of particular phases of the scientific method being illustrated.

As a note for the test developer, as he/she proceeds though the later steps, specification of stimulus and response attributes, this general description may have to be reworked several times.

The second component is the specification of a sample test item, including the directions to the student about how to respond. Popham (1978a) feels there are two reasons for providing a sample item a the second component: It serves as an illustration for individuals unable to read the detailed descriptions (because of time involved) and, more importantly, it provides format cues for the item writers. It specifies the preferred form in which the items can be constructed. Popham (1978a) also suggests that the correct answer not be identified for the sample item; the complete specification should be considered before the reader can adequately assess degree of correctness of the sample item.

The third step in the development of a domain specification by far the most difficult; here the attributes of the stimulus materials are specified, along with delimitations on possible stimuli. In other words, there must be an extensive, i.e, as complete as possible, description of what stimuli can constitute a test item. According to Popham (1978a):

> In the stimulus attributes section of the test specifications we must set down all the really influential factors that constrain the composition of a set of test items.

Further, Popham (1978a) notes:

> The general rule is that the test specifier has to spell out all of the critical and controlling dimensions which will permit someone to create a set of test items that will, without exception, be viewed as congruent with the constraints set forth in the specifications.

For tests that depend on subject matter content, which most criterion-referenced tests do (although one could envision CRTs in the affective domain), Popham suggests that one of three techniques be utilized in specifying content that can be used for stimuli (in the test item). One technique is to spell out rules or algorithms which are used in generating and delimiting the content. Item generation rules or item forms analysis

(Hively et al., 1973) is an example of such an approach. As mentioned before, we see this as being an "idealized" situation likely to occur for only structured subject areas. However, if item forms analysis is possible, it is at this point that Popham's work and the work of Hively and his associates unites. A second technique for delimiting and specifying content is to list all the content that might be included. This would seem to hold relevance for even fewer situations than the first technique. The final technique, which holds the most relevance for situations the test practitioner is likely to encounter, is to try as carefully as possible to isolate and describe the defining attributes of all eligible content for the test. Popham says that if even this isn't possible, some examples of acceptable and unacceptable content is better than nothing.

In consideration about what does and what does not constitute relevant stimulus attributes, Popham suggests always using the following reminder:

> What are the absolutely indispensable elements that items writers must consider in producing test items?

Further, in deciding about what rules for inclusion or exclusion of stimuli need to be specified, preparation of some trial items should be helpful in making decisions.

In sum, the specification of stimulus attributes constitutes the most critical step in the development of a domain specification. The rules for generating the items for the domain, and ultimately the test itself, must be defined here. Therefore, the stimulus attributes section should both specify the content on which the items are generated and also describe the "directions to respond" that the student is to receive.

The fourth step, specification of the response attributes, is a little easier than specification of the stimulus attributes. This is because only two possible types of responses can be made by the examinee. He/she can either select from a set of response options for a test question or he/she can construct a response. This section should specify the rules/criteria upon which both sorts (if both are used), or a single sort of response type is to be treated.

If an examinee has to select a response, rules must be provided for determining the nature of the correct and the incorrect responses. In other words, when given to an item writer, the writer should be able to generate the correct response and the incorrect response(s) directly from the response attribute section. Popham (1978a) suggests that identification of wrong answer options for a test item can usually come about by considering the various ways in which the examinee "goes wrong."

If the respondent is asked to construct his/her own response, the task of specifying response attributes becomes even more difficult. The test developer must try to explicate the criteria that should be used to judge how adequate an examinee's response is. Popham again suggests that creation of some sample trial responses should help in this process. Finally, in reference to constructed responses, Popham warns of the use of "hedging phrases." He uses examples such as "responses must be appropriate to the context of the stimulus" or "answers should be reasonable outgrowths of the materials provided." Without further defining "appropriate" and "reasonable", such explications of response attributes has accomplished nothing. The explication must be specific, so that, based upon the response attributes, one can ascertain the "fit" of the student's response to the specifications.

The specification supplement is just that; a supplement that might contain information on the stimulus attributes and/or response attributes section(s) that would have made the respective sections too long. In other words, contained here would be, for instance, content listings that should be listed some place, but aren't critical for the stimulus attributes section.

Following through these series of steps should help the test developer come up with a quite well-developed set of domain specifications. Unlike the situation when alternate approaches such as item forms analysis are used, one can never be certain that all sources of ambiguity have been removed from a domain specification. However, it should be quite clear to the reader at this point that the procedure just described is relevant for a wide variety of subject domains, and from a practical point of view, one should be able to live with the ambiguity because of the procedure's great flexibility.

No matter how much care is taken in preparing domain specifications, they must be carefully reviewed by other content specialists and individuals who will use them (for example, item writers and teachers). On the next two pages is a draft copy of a domain specification review form. It is assumed in the review form that a domain specification is divided into four sections: (1) Skill, (2) Sample Directions and Test Items, (3) Content Domain, and (4) Characteristics of Answer Choices and Scoring. (Clearly, the four sections correspond to those proposed by Popham but we have used new section labels to facilitate communication with domain specification writers.)

Our usual procedure is to separate domain specification writers into work groups of three or four. Their task is to produce draft copies of domain specifications. (The overall content of each domain specification is specified in a "test blueprint" or "content guide" which must be prepared and approved before the writing of any domain specifications can begin.) The draft copies are then critiqued by at least one other work group (and more if there is time and money to do so) using the review form to guide the direction of the critique. Once a domain specification is reviewed, the writing group and the review group(s) meet to discuss the review group's critique. Following this meeting, appropriate revisions can be made. It is usuallv desirable to have the revised domain specifications reviewed again, sometimes by a larger and more diverse group of individuals.

Domain Specification
Review Form

Domain Specification:________ Reviewer:______________________ Date:_______________

Please read the domain specification carefully. Next answer the eight questions below. Staple (or clip) your copy of the domain specification to this review form after you have answered the questions.

Thank you for your comments.

SKILL

1. Does the "skill section" provide sufficient details to give a reader an indication of the behaviors defined by the domain specifications? (Circle one)

   (a) Yes
   (b) Yes, with reservations
   Please explain:

   (c) No

2. How might the "skill section" be revised to improve its clarity?

Sample Directions and Test Item

3. Can the test directions be revised to improve their clarity? (Circle One)

   (a) Yes, (please write your comments on the domain specification)
   (b) No, they are clearly written

4. Do you feel the item format is the best to ensure that examinee answers will measure the behaviors defined by the general description? (Circle One)

   (a) Yes
   (b) Yes, with reservations
       Please explain:

   (c) No. Which item type would be best?

5. Will the sample test item provide a "good model" for item writers preparing items to measure the domain specification? (For example, does the item include the desired number of answer choices, and is the vocabulary appropriate for the intended group of examinees?) (Circle One)

   (a) Yes
   (b) Yes, with reservations
       Please explain:

   (c) No

CONTENT DOMAIN

6. Do you have any suggestions for revising and/or extending the content defined by the domain specification? Please write comments on your copy of the domain specification (Your suggestions could include: (a) deletions of specific content, (b) additions to the content, (c) rewrites for clarification).

CHARACTERISTICS OF ANSWER CHOICES AND SCORING

7. Do you have any suggestions for revising and/or extending the characteristics of possible answers? Please write comments on your copy of the domain specification.

8. (For essay questions only.) Do you have any suggestions for improving the scoring of test items? Please write comments on your copy of the domain specification.

2.4 Examples of Domain Specifications

The examples of domain specifications offered in this section come from many sources. The first two examples are from Popham (1978a) and are included here with his permission. They are direct applications of the steps discussed in section 2.3. The third example was prepared by Millman and Craig (1977) and reproduced here with the senior author's permission. Although the domain specification is organized in a fashion different from the plan offered in the last section, it is an excellent example of a domain specification. The next set of examples were prepared by Jerry George and his staff at the Glendale Union High School District in Arizona. They are included in our materials with their permission. These sample domain specifications are only a few of more than a hundred they have prepared in the last two years in four subject areas. Again, the format of the domain specifications is different from Popham's most recent recommendations advanced in section 2.3, but clearly the chosen format represents an attempt by the authors to clarify the relevant domain of behaviors defined by the objectives defining their high school curricula. (They prefer the term "criterion-referenced test model" to "domain specification." The first term was introduced by Popham a number of years ago.) The remaining seven examples are second or third drafts of domain specifications prepared at workshops offered by the authors of this Practitioner's Guidebook. The domain specifications are from several content areas and presented in several different formats.

Example 1 (Popham, 1978, pp. 129-131)

An illustrative set of criterion-referenced test specifications: applying concepts of United States foreign policy

General description

Given a description of a fictitious international situation in which the United States may wish to act, and the name of an American foreign policy document or pronouncement, the students will select from a list of alternatives the course of action that would most likely follow from the given document or pronouncement.

Sample item

Directions: Read each fictitious example below. Decide what action the United States would most likely take based on the given foreign policy document. Write the letter of the action on your answer sheet.

> Some Russian agents have become members of the Christian Democratic Party in Chile. The party attacked the president's house and arrested him. The Russian agents set themselves up as president and vice-president of Chile. Chile then asked to become an "affiliated republic" of the USSR.
>
> Based on the *Monroe Doctrine,* what will the United States do?
>
> *a.* Ignore the new status of Chile.
> *b.* Warn Russia that its influence is to be withdrawn from Chile.
> *c.* Refuse to recognize the new government of Chile because it came to power illegally.
> *d.* Send arms to all groups in the country that swear to oppose communism.

Stimulus attributes

1. The fictitious passage will consist of 500 words or less followed by the name of a foreign policy pronouncement or document inserted into the question, "Based on the ____________, what will the United States do?

2. The policy named in the stimulus passage will be a document or pronouncement selected from the specification supplement.

3. Each passage will consist of two parts: (a) a background description of an action taken by a foreign nation and (b) a statement of the action to which the foreign policy document or pronouncement is to be applied.

a. The background statement will be analogous to an historical situation that either preceded the issuance of the cited document or pronouncement or for which the document or pronouncement was used. For example, the Monroe Doctrine was drawn up in response to European designs on American nations that were attempting to establish independence. An analogous case today might describe a European country attempting to encroach on the sovereignty of an American country.

b. The statement of an action will describe an action taken by a real foreign nation that conforms to one of the following categories:

(*1*) Initiation of an international conflict.

(*2*) Initiation of a civil conflict. This may include coups, revolutions, riots, protest marches, civil war, or a parliamentary crisis.

(*3*) Initiation of an international relationship. This may include trade negotiations, friendship pacts, military alliances, and all classes of treaties.

(*4*) Appeal for foreign aid to meet economic or military needs.

(*5*) Development and stockpiling of military weapons.

4. All statements in the passage will refer to specific nations and events. Descriptions such as, "A nation is at war with another country," are not acceptable.

5. When the document or pronouncement mentioned in the stimulus passage is tied to a particular geographical region, countries named in the passage must belong to that region.

6. Passages will be written at no higher than the seventh-grade reading level.

Response attributes

1. Students will be asked to mark the letter of one of four given response alternatives consisting of the correct response and three distractors. Each alternative will possess the following characteristics:

a. Describe a specific course of action that refers to the people, nations, and actions in the stimulus passage.

b. Be brief phrases written to complete the understood subject, "The United States will . . ."

2. Distractors (wrong answers) will be written to meet these additional criteria:

a. Each distractor will describe an action derived from a different document or pronouncement selected from the specification supplement.

b. Documents or pronouncements from which identical courses of action may be derived will not be used.

c. The decision for the United States not to act may be used as a course of action when it is based on a document or pronouncement.

3. The correct response will be the course of action that is governed by the principles described in the document or pronouncement named in the stimulus passage.

Specification supplement: eligible policy documents and pronouncements

The following list of foreign policy pronouncements and documents was selected from Thomas Brockway, *Basic Documents in United States Foreign Policy* (Princeton, N.J., D. Van Nostrand, 1968). The items selected were chosen on the basis of their generalizability and potential application to current events. The list appears in chronological order.

1. The Declaration of Independence
2. Washington's Farewell Address
3. The Monroe Doctrine
4. Webster on Revolutions Abroad
5. Open Door in China
6. The Platt Amendment
7. Roosevelt Corollary of the Monroe Doctrine
8. The Fourteen Points
9. The Washington Conference
10. The Japanese Exclusion Act
11. The Kellogg-Briand Pact
12. The Stimson Doctrine
13. Roosevelt's Quarantine Speech
14. The Atlantic Charter
15. The Connally Resolution
16. The Yalta Agreements
17. The Potsdam Agreement
18. United States Proposals for the International Control of Atomic Power
19. The Truman Doctrine
20. The Marshall Plan
21. The Point Four Program
22. The North Atlantic Treaty
23. American-Japanese Defense Pact
24. Atoms for Peace: Eisenhower's Proposal to the United Nations
25. The Formosa Resolution
26. The Eisenhower Doctrine
27. Alliance for Progress
28. Kennedy's Grand Design
29. Treaty on the Peaceful Uses of Outer Space

BEST COPY AVAILABLE

BEST COPY AVAILABLE

Example 2 (Popham, 1978, pp. 132-134)

An illustrative set of criterion-referenced test specifications: job interview procedures

General description

Having read a description of a job interview in which the applicant may make one of several specified types of errors in appearance, conduct, or preparation, the student will select the error made or indicate that no error was made.

Sample item

Directions: Read the description of each job interview below. If the applicant makes an error in interview behavior, mark the letter of the response alternative that matches the error described. If no error was made, mark "e."

> Anita arrives five minutes early for an interview for a trainee job in floral design and sales. She wears a white dress with long, full sleeves and shoes with high heels. She brings a portfolio of her work as a design major in high school and briefly points out the designs she feels are most closely related to floristry. She answers the interviewer's questions in a brief, courteous manner and indicates her willingness to perform all aspects of the florists' trade, including scrubbing floors, washing buckets, and disposing of spoiled flowers.
>
> What is Anita's error?
>
> *a.* lack of punctuality
> *b.* inappropriate dress
> *c.* irrelevant materials presented
> *d.* inappropriate attitude
> *e.* no error was made

Stimulus attributes

1. Each item will consist of a fictitious description of 100 words or less dealing with a named person's job interview, followed by that person's name inserted into the question, "What is ________ ________'s error?"

2. The description will include the type of job being applied for and illustrations of at least four of the following behavioral factors that may influence an impression of an applicant:

a. Punctuality—arrival at or within a reasonable time before the specified interview time. Arrival after the specified time, or arrival more than ½ hour early will be considered lack of punctuality, as both may inconvenience the interviewer.

BEST COPY AVAILABLE

b. Appropriateness of dress—dress which is neat, clean, and practical for the type of job being applied for. If one expects that an interview may include a demonstration of skills, one's clothing must not interfere with such a demonstration. Extremes such as very high heels, low cut dresses, very tight pants, etc., are almost always inappropriate. Appropriateness of dress also includes such personal grooming items as length of fingernails, length and style of hair, etc., which are inappropriate only if they are likely to interfere with the work involved in the job being applied for (e.g., long fingernails on a secretarial applicant).

c. General courtesy—pleasantness and politeness to all individuals encountered before, during, and after the interview.

d. Frankness—honesty and directness in answer to personal or experience-related questions. False answers, misleading answers, attempts to change the subject, or attempts to rationalize answers will be considered lack of frankness.

e. Careful thought to answers—brief, clear, well-thought-out answers to problems posed by interviewer. Excessive wordiness, self-contradiction, disorganized answers, and answers that do nothing more than reiterate the problem will be considered evidence of lack of careful thought to answers.

f. Appropriateness of attitude—interest and enthusiasm displayed toward all aspects of job, but without pushiness or opinionatedness. Interest and enthusiasm may be indicated by simply stating their presence (e.g., "John appears very interested in the techniques demonstrated") or by a direct or indirect quotation on the part of the applicant expressing enthusiasm or interest (e.g., "Of course I don't mind emptying buckets. I want to learn all about the business."). Pushiness and opinionatedness may be indicated by attempts to tell the interviewer how the business should be run, boasting about superiority of knowledge or ability (as opposed to offering to demonstrate ability), sarcastic comments, attempts to bully interviewer, and similar actions. General lack of enthusiasm (indicated by description or quotation), complaints about specific aspects of the job, or the presence of any of the indications of pushiness or opinionatedness will be considered inappropriate attitude.

g. Relevance of materials presented—direct and obvious relationship to job being applied for of any education- or experience-related materials brought to interview. Examples of appropriate materials are a typing award for a secretarial applicant, or a portfolio of works from a high school design course for an applicant in any art- or design-related field. Examples of inappropriate materials are a tennis award for an engineering applicant, or a record of offices held in high school for a janitorial applicant. The relevance or irrelevance of such materials may be made more obvious by describing the applicant's mode of presentation (e.g., "She brings a portfolio of her work as a design major in high school and briefly points out the designs she feels are most closely related to floristry.") or by indicating the purpose of the applicant in bringing the material (e.g., "John, who is applying for a job as an engineer, brings a letter of recommendation from his previous employer (who runs a hamburger stand) to show his reliability and industriousness.")

h. Specific and realistic goals—applicants' ability to explain their purpose for applying for the job (to start a career, earn money for college, etc.) and what working conditions, salary, and rate of advancement they expect. Inability to answer specific questions dealing with these issues (e.g., "What salary do you expect?" "I don't know. What did you plan to pay?") or working conditions, salary, or advancement expectations that are exceptionally high or low for the job being applied for (e.g., plans to be vice-president of company within two years of being hired as a secretary, or asking only $2.50 per/hour for work requiring a graduate degree or highly specialized training), will be considered lack of specific and realistic goals.

3. The interview description may illustrate completely correct behavior, or one of the behavioral factors illustrated may exemplify erroneous behavior, whereas the rest of the description exemplifies correct behavior. No more than 20 percent of the test items will exemplify completely correct behavior.

4. The description may include direct quotation of the interviewer and/or the interviewee, as well as description of their actions and conversation.

5. If several descriptions are used in a test, the names given to interviewers will be evenly divided between male and female, and will include some named characteristic of the most common ethnic groups in the population to be tested. The name to be used with a given job will be chosen at random so that discrimination cannot be made on the basis of sex or ethnic group.

6. The readability of the descriptions will be no higher than tenth-grade level.

Response attributes

1. The students will mark on their answer sheets the letter that corresponds to the error made by the job applicant (if any) or the statement that "no error was made."

2. There will be five alternatives, consisting of the correct response and four distractors. The options will include the response "no error was made" along with four of the following behavior factors: lack of punctuality, inappropriate dress, lack of general courtesy, lack of frankness, lack of careful thought to answers, inappropriate attitude, irrelevant materials presented, and lack of specific and realistic goals. The four behavioral factors chosen will correspond to four of the factors illustrated in the interview description and will include that factor (if any) in which an error is illustrated.

3. The correct response will be that alternative that correctly names the error illustrated in the description of the interview description, or, in the event that no error was illustrated, that alternative that states "no error was made."

BEST COPY AVAILABLE

Example 3 (Millman & Craig, 1977)

EXAMPLE OF A DOMAIN SPECIFICATION[1]

UNIT PRICING

Objective:

Identify the package having the lowest unit price, given different sizes of the same brand product and their cost. (Similar to performance indicator 4E1.)

Rationale:

Retail items are an important component of every consumer's budget, and an understanding of unit pricing is essential to economic buying habits.

---

[1] From Millman, J., & Craig, M. M. Rhode Island's educational performance indicators and items: An independent evaluation and feasibility report. Final Report. June 1978. (Reproduced with permission from the senior author.)

Sample Items:

1. The unit price labels for three packages of Boundless paper towels, each of a different size, are shown below.

Boundless Towels	2 ply
UNIT PRICE	RETAIL PRICE
7.7¢ per sq. yd.	86¢ 100 sq. ft. 50 sheets

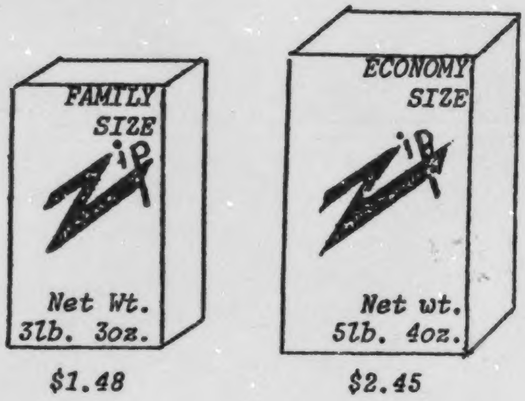
Boundless Towels	2 ply
UNIT PRICE	RETAIL PRICE
7.1¢ per sq. yd.	$1.19 150 sq. ft. 75 sheets

Boundless Towels	2 ply
UNIT PRICE	RETAIL PRICE
8.1¢ per sq. yd.	$2.25 250 sq. ft. 125 sheets

Which package is the most economical?

A. 50 sheet size
B. 75 sheet size
C. 125 sheet size
D. More information is needed to answer the question.

2. Two packages of Zip detergent are shown below.



Which package has the lower unit price?

A. Family size
B. Economy size
C. They have the same unit price.

Questions

The test items will be about familiar objects found in a grocery or similar retail store.

Within a given item, all packages will be of the same brand and product (and thus may be assumed to have the same ingredients).

Relevant information on unit and price will be presented in one of three forms:

a) unit pricing labels,
b) pictures or drawings of product packages, or
c) written text on unit and price.

Information on "coupons inside package" or "cents off the marked price" will not be included.

In each question, two to four packages will be identified.

The question will ask which package either has:

a) "the lowest unit price" or is
b) "the most economical".

Note: "Best buy" language will not be used since considerations other than lowest unit cost or economy enter into the determination of best buy.

Prices will be shown using the $ symbol for amounts of $1.00 or more; the ¢ symbol will be used otherwise.

Across items, there will be no relation between the size of the package and its unit price.

Options

Responses will be presented in multiple-choice format.

There will be only one correct answer per question.

When only two packages are shown, "They have the same unit price" will be used as one of the response options.

The option, "More information is needed to answer the question", can be used.

Units of Quantity

The units may be of number, weight, length, area, or volume.

More than one unit may be shown on any one package. A roll of toilet tissue, for example, may indicate both the number of sheets and the area.

If the units are not comparable across products, the student is to answer that more information is needed to answer the question. A brand of shampoo sold in both liquid and paste form, for example, may be expressed in noncomparable units.

Mathematics Involved

If conversions are required to make the units comparable, then the conversion factors will be provided as part of the problem.

Prices will be chosen so that the most economical item will be at least a full penny less expensive than other items, whether or not students round off their calculations.

Students will not have to compute areas.

Example 4 (Glendale Union High School District)

PROGRAM	Language Arts	SUBDIVISION	I.A.1.a. - g.
COURSE	English 1-2	SKILL/CONCEPT	Parts of Speech
TEST	Pre/Post Test	BEHAVIORAL LEVEL	Application

RATIONALE

Parts of speech are taught, not as ends in themselves, but as tools for the improvement of oral and written communications. Instructions should help students see how parts of speech work together to form meaningful language structures.

CONTENT LIMITS

1. A noun is the name of a person, place, thing, or idea (actor, city, automobile, kindness).
2. A pronoun is a word which can be substituted for a noun.
3. Personal pronouns may refer to the person speaking, the person spoken to, or the person spoken of (I--me, you, he--him, she--her, we--us, they--them).
4. Indefinite pronouns do not refer to any specific persons or things (everybody, everyone, somebody, someone, nobody, no one, anybody, anyone).
5. A verb is a word that shows action or state of being (rush, bite, is, are).
6. A verb which is made up of more than one word is called a verb phrase (is leaving, was helping).
7. An adjective is a word that describes, limits, or modifies a noun or pronoun (a walnut desk, a cloudy day, he is remarkable).
8. An adverb is a word that describes, limits, or modifies a verb, adjective, or another adverb (came quickly, an especially fine paper, played fairly well).
9. A preposition is a word which shows a relationship between its object and some other word in a sentence. (We flew above the clouds. They lived around the corner. She is at home.)
10. A conjunction is a word which joins words or groups of words (and, but, or nor, for, yet).
11. Correlative conjunctions are used in pairs with other words dividing them (both--and, either--or, neither--nor, not only--but also, whether--or).

ITEM FORMAT

Format is one simple or compound sentence with five words lettered and underlined, followed by two or three questions asking for identification of underlined words.

(continued)

ITEM FORMAT (CONTINUED)

Item Restrictions

1. The following will not be tested or used as distractors:
    - verbals
    - a, an, and the
    - prepositions ending in ing (during, concerning, etc.)
    - phrasal prepositions (according to)
    - adverbs which also function as adjectives or prepositions (deep, up, etc.)
    - pronouns other than the examples shown in the content limits
2. Sentences developed for testing parts of speech will not contain infinitive phrases.

RESPONSE DESCRIPTION

The student will demonstrate application of rules for distinguishing parts of speech by identifying parts of speech in simple and compound sentences.

CRITERIA

The student will correctly identify the underlined part of speech.

DIRECTIONS TO THE STUDENT

Parts of Speech. Read the sentence and answer each of the questions below it by blackening the lettered space on your answer sheet.

ITEM

A B C D E
The suspect led police on a wild, high-speed chase through the city.

1. Which of the underlined words is a noun? (D)
2. Which one is a preposition? (E)

Example 5 (Glendale Union High School District)

PROGRAM Language Arts
SUBDIVISION I.B.1.a.
COURSE English 1-2
SKILL/CONCEPT End Marks (MODEL A)
TEST Pre/Post Test
BEHAVIORAL LEVEL Application

RATIONALE

Question marks and periods at the ends of sentences are taught to help students achieve clarity in written communications.

CONTENT LIMITS

1. A period is used at the end of a declarative sentence. (I spoke to them.)
2. A question mark is used at the end of an interrogative sentence. (Did you speak to them?)

ITEM FORMAT

Format is a set of four sentences, one of which contains an error in the use of the period or question mark. Each set of four sentences will contain

1. One or more direct statements
2. One or more direct questions
3. One indirect or false question

RESPONSE DESCRIPTION

The student will demonstrate application of the rules for using periods and question marks at the ends of sentences by identifying errors in the use of end mark punctuation.

CRITERIA

The student will identify the sentence which is improperly punctuated.

DIRECTIONS TO THE STUDENT

Using Periods and Question Marks. Find the sentence which contains the error. Blacken the lettered space on your answer sheet.

ITEM

1. A. The research papers will be due on March 13.

   B. He wants to know whether your parents will attend the concert?

   C. Why do you think your teacher made such strict rules?

   D. Mr. Carson will be in Europe during July and August.

Example 6 (Glendale Union High School District)

PROGRAM Reading

SUBDIVISION II.C.1.b.

COURSE Modern Reading Techniques

SKILL/CONCEPT Paragraph Meaning--Inferred Main Idea Int. & Jr.

TEST

BEHAVIORAL LEVEL Synthesis

RESPONSE DESCRIPTION

Given ten paragraphs with the stated main idea omitted, the learner will demonstrate synthesis of the skill of finding inferred main ideas by choosing the correct statement of inferred idea from four possible choices.

CONTENT LIMITS

The reader is advised to find inference in the following manner:

1. Look at the details to see if they have something in common.
2. Make a guess about the intended meaning.
3. Review the clues or ideas to see if they support your guess.

An inferred main idea, then, may be defined as the idea or fact that is not stated but is drawn from stated facts or ideas. Inference involves a process of inductive reasoning.

ITEM FORMAT

The learner will be presented with ten paragraphs followed by four answers which will state, or complete, an inferred main idea, one of which will be correct, and three of which will be distractors. The paragraphs are graded in level of difficulty from grade 3 through 8. (Questions 1-5 are intermediate, grades 3-5. Questions 6-10 are junior, grades 6-8.)

CRITERIA

The learner must select the correct inferred main idea with 60% proficiency.

PROGRAM Reading SUBDIVISION II.C.1.b.
COURSE Modern Reading Techniques SKILL/CONCEPT Paragraph Meaning--Inferred Main Idea Int. & Jr.
TEST BEHAVIORAL LEVEL Synthesis

DIRECTIONS

Paragraph Meaning--Inferred Main Idea

You will be asked to draw a conclusion about what is really meant in each of the following passages. After carefully reading each paragraph, on the answer sheet blacken the space of the letter which best states the inferred meaning.

ITEM

EXAMPLE:

The house was run-down. After twelve years it still was not painted. There was no porch; crude wooden steps led up to the warped front door. The outside light, hanging down by its cord, swung to-and-fro with the night breeze. The house was unfinished on the inside too. The ceiling was only plasterboard haphazardly nailed in place. Paint and plaster were cracking and flaking onto the floor from the walls.

(a) The house had not been painted.
(b) the inside of the house had plasterboard ceilings.
(c) The paint flaked onto the floor.
(d) The house was in disrepair.

The correct answer is "d."

---

1. I'm thinking, I'm thinking
So leave me alone.
I don't need your help.
I'll do fine on my own.

I have a few problems
I have to work out,
Which cannot be done
If you stand there and shout.

"We need you for baseball,
So come right away."
I'll come when I feel
I am ready to play.

Please stop making faces.
It won't help to grown.
I'm thinking, I'm thinking,
So leave me alone.

The writer ____________________.

(a) dislikes everybody all the time
(b) never thinks by himself
(c) sometimes likes to be alone to think
(d) is not liked by other people

Example 7 (Glendale Union High School District)

COURSE(S) Algebra 1-2

SUBDIVISION V.E.

BEHAVIORAL LEVEL Application

SKILL/CONCEPT Factoring Polynomials With More Than Three Terms

RESPONSE DESCRIPTION:

Factor a polynomial of more than three terms.

CONTENT LIMIT:

A polynomial of four terms of the form $a^2 + 2ab + b^2 - c^2$ where c is an integer between 0 and 5, inclusive.

ITEM FORMAT:

See I.A.1. for a, b, c

- d) One problem
- e) One wrong answer (b) will be $(a + c)(a - c)(2ab + b^2)$
- f) One wrong answer (c) will be $b(a + c)(a - c)(2a + b)$
- g) One wrong answer (d) will be $(a + b - c)^2$

CRITERIA:

Select correct answer

DIRECTIONS:

ITEM:

Factor Completely:

$a^2 + 6ab + 9b^2 - 25$

- * a) $(a + 3b + 5)(a + 3b - 5)$
- b) $(a + 5)(a - 5)(6ab + 9b^2)$
- c) $3b(a + 5)(a - 5)(2a + 3b)$
- d) $(a + 3b - 5)^2$
- e) None of the above

Example 8[1]

SKILL: The student will identify the tone or emotion expressed in a paragraph.

SAMPLE DIRECTIONS AND TEST ITEM:

Directions: Read the paragraph. Underline the best word to complete the sentence.

Jimmy had been playing at the beach all day. It was time to go home. Jimmy sat down in the back seat of the car. He could hardly keep his eyes open.

Jimmy felt __________________.

A. afraid B. friendly C. tired D. kind

CONTENT DOMAIN:

1. The paragraph will contain situations which are familiar to the students being tested.
2. The paragraph will contain no less than three and no more than six sentences. The readability level will be no higher than Second Reader.
3. The tones or emotions expressed will be from the following list:

sad
tired
happy
kind
mad
scared
lucky
excited
angry
friendly
smart
proud

RESPONSE MODE:

1. Responses will be one word in length.
2. The items will contain one correct and three incorrect responses.
3. Distractors are to be words describing a feeling and may be taken from the list above.
4. Avoid using reasonable answers as distractors (i.e., in the sample item, "mad" would not be a good choice for a distractor—Jimmy could feel mad about leaving the beach).

---

[1] An example of a domain specification from the reading area. (The authors are grateful to Marlene Teichert of Educational Progress for the example.)

Example 9[1]

Content: Reading
Strand: Comprehension
Level: 2

SKILL

A student will be able to identify the main idea of a paragraph by choosing the best title.

SAMPLE TEST DIRECTIONS AND TEST ITEM

Test Directions

Read each paragraph and choose the best title. Circle the letter beside your answer.

Test Item

The second grade went on a class trip. They saw airplanes and jets. A man told them how to buy an airline ticket. They saw the pilot's cockpit.

a. Meeting a pilot.
b. Buying a ticket.
c. A trip to the airport.

CONTENT DOMAIN

1. Sentences should have no less than 3 words, and no more than 10.
2. Each paragraph should have no less than 3 sentences and no more than 7.
3. Compound and simple sentences should be included.
4. Readability should be approximately 2.5.
5. The paragraphs should include both experience and interest-oriented subject matter.

CHARACTERISTICS OF ANSWER CHOICES AND SCORING

1. There should be three titles to choose from. The correct title is the main idea of the paragraph.
2. Distractors should contain smaller details from the paragraph.

---

[1] An initial draft of the content included in this domain specification was prepared by Liz Jerrett and Nancy Cole.

Example 10[1]

Content: Reading
Strand: Structural Analysis
Level: 3

SKILL

The student will identify the meaning of a word consisting of a root word and a prefix.

SAMPLE TEST DIRECTIONS AND TEST ITEM

Test Directions

Read the word and the two definitions that follow it. Choose the correct meaning of the word and place the letter (A) or (B) on the line in front of the word.

Test Item

___1. disappear	(A) to appear again	(B) to drop out of sight
___2. exterior	(A) outside	(B) inside
___3. incapable	(A) can do	(B) can't do

CONTENT DOMAIN

1. The stimulus words will contain the following prefixes:

   un re in dis ex

2. The words are to be at a vocabulary level no higher than level four.
3. The words are not to be included in the context of a sentence.
4. See attached list of words for suggested content.

CHARACTERISTICS OF ANSWER CHOICES AND SCORING

1. The student will write the letter of the correct response on the line provided.
2. There will be two choices, the correct response and one distractor.
3. The distractor will contain a meaning for the root word without the prefix or the meaning of the root word with a different prefix.

---

[1] An initial draft of the content included in this domain specification was prepared by Marlene Teichert.

Suggested List of Words

un	in	re	dis	ex
uneven	insincere	remind	disown	exclude
unclean	inhuman	reform	disconnect	exclaim
unfold	insight	rename	discover	exhale
untie	incapable	regain	disband	exit
unreal	informal	rejoin	disloyal	expand
unsafe	inability	replant	displease	expel
untrue	inclose	retold	dishonor	expire
unfit	indent	recall	discount	explain
uneasy	inland	reopen	dismount	explore
unhappy	indoor	renew	disarm	extend
unpack	incomplete	reread	disorder	exterior
unload	intake	refill	disable	

Example 11[1]

Content: Mathematics
Strand: Fractions
Level: 4

SKILL

The student will be able to multiply fractions.

SAMPLE TEST DIRECTIONS AND TEST ITEM

Test Directions

Circle the answer to the question below:

Test Item

$\frac{1}{3} \times \frac{3}{4} =$

a. $\frac{3}{7}$ b. $\frac{4}{7}$ c. $\frac{1}{4}$ d. $\frac{4}{9}$ e. $\frac{9}{4}$

CONTENT DOMAIN

1. Fractions will be written using the horizontal bar $(\frac{a}{b})$.
2. Limit to fractions less than one with single digit denominators.
3. The numerators and denominators of each fraction in an item stem will have no factor in common other than one.
4. In the item form $\frac{a}{b} \times \frac{c}{d}$ or $\frac{c}{d} \times \frac{a}{b}$, *a* and *d* will have no common factor except one, and each of the following cases will be included in the items:

   a. *b* is a multiple of *c*, or *c* is a multiple of *b*;
   b. *b* and *c* are equal;
   c. *b* and *c* share a common factor other than one;
   d. *b* and *c* share no common factor except one.

---

[1] An initial draft of the content included in this domain specification was prepared by a group of teachers working with the Accountability Renewal Model Project in Texas.

CHARACTERISTICS OF ANSWER CHOICES AND SCORING

1. Each item will contain five answer choices, only one of which is correct.

2. "None of these" will not be used as a answer choice.

3. Distractors will represent the most frequent student errors.

4. Distractors will include errors such as:

   a. multiplication of numerators and addition of denominators;
   b. "cross multiplication" (numerator x denominator) either end up;
   c. addition of numerators and denominators.

5. Equivalent forms of the correct answer will not be used in a set of answer choices.

Example 12[1]

Content:	Mathematics
Strand:	Life Skills
Level:	7

SKILL

Student will use reference units of weight/mass, length, area, volume, temperature, time, and money to estimate and determine measures, both metric and customary.

SAMPLE TEST DIRECTIONS AND TEST ITEM

Test Directions

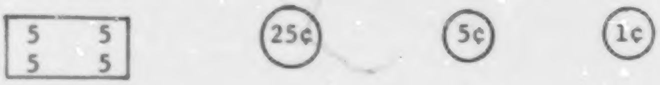
Circle your answers to the questions below:

Test Items

1. The distance from Fort Worth to Austin would be measured in ________.

    a. kilometers
    b. kiloliters
    c. kilograms
    d. liters
    e. grams

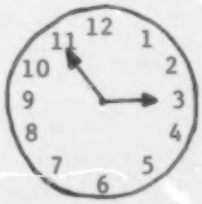
2.



Find the value of money shown above.

    a. $5.26
    b. $5.06
    c. $25.51
    d. $5.31
    e. $1.40

3.



What is the correct time?

    a. 11:15
    b. 2:50
    c. 2:55
    d. 3:55
    3. 3:50

---

[1] An initial draft of the content included in this domain specification was prepared by a group of teachers working with the Accountability Renewal Model Project in Texas.

CONTENT DOMAIN

1. Limit the unit measurements to the following:

   kilometer, meter, centimeter, millimeter, mile, yard, foot, inch, the square unit of the previous listed, the cubic unit of the previous listed, liter, milliliter, kilogram, gram, gallon, quart, ton, pound, ounce, hour, minute, second, degree Fahrenheit, degree Celsius.

2. The following types of items will be included:

   a. Items which reference a real life object familiar to students, and a characteristic of the object to be measured. The student will select the appropriate unit to measure the object. Items will include measurement of weight/mass and capacity; area and volume; and length.

   b. Items which deal with money. The student will select the total amount of money or the appropriate collection of bills and coins. There will be no more than 9 or fewer than 3 bills and/or coins pictured in the stems or response choices.

   c. Items which deal with time, length, and temperature. The student will be presented a picture and will give the appropriate time, length of object, or temperature. Time will be measured to the nearest minute. Temperature will be measured to the nearest degree.

3. Conversions will not be required.

4. Items will not test knowledge of abbreviations.

CHARACTERISTICS OF ANSWER CHOICES AND SCORING

1. Each item will contain five answer choices, only one of which is correct.
2. "None of these" will not be used as an answer choice.
3. Distractors will represent the most frequent student errors.
4. More than one unit that measures the same characteristic will not be used in the answer choices for an item described in content 2a.
5. Answer choices will include the unit of measurement, where appropriate.

Example 13[1]

Content: Mathematics
Strand: Life Skills
Level: 7

SKILL

A student will identify the page (or pages) from a newspaper index on which information related to a given topic can be found.

SAMPLE TEST DIRECTIONS AND TEST ITEMS

Test Directions

Please read the newspaper index below and answer the questions. Circle the letter beside your answer to each question.

Test Items

The Austin Record News

Amusements	E5-7	Horoscope	F1
Classified Ads	F4-6	Personalities	B5
Comics	C11	Sports	C1-8
Editorials	A14	TV Logs	E9
Financial	D4,5	Weather	F3

1. Where would you find information about a person born under the sign of Scorpio?

   (a) F4-6 (b) B5 (c) F3 (d) F1

2. Where would you find the necessary qualifications for available jobs?

   (a) E9 (b) F4-6 (c) B5 (d) F3

3. If you wanted to find the standings in the National Football League, where would you look?

   (a) C1-8 (b) E5-7 (c) C11 (d) E9

4. Where would you read a person's opinion on a current local political matter?

   (a) A14 (b) C11 (c) D4,5 (d) C1-8

---

[1] An initial draft of the content included in this domain specification was prepared by a group of teachers working the Accountability Renewal Model Project in Texas.

CONTENT DOMAIN

1. A newspaper index will be reproduced containing both a section letter and page(s) for each topic.

2. The newspaper index will contain no more than ten topics listed in one or two columns.

3. The ten (maximum) topics selected will be from the following list:

   Amusements
   Classified Ads
   Comics
   Crosswords
   Deaths
   Editorials
   Financial
   Food
   Horoscope
   Sports
   TV Log
   Weather
   Personalities

4. Test items will relate to the topics listed but will not name the topics.

CHARACTERISTICS OF ANSWER CHOICES AND SCORING

1. There will be one correct and three incorrect answers for each test item.

2. Avoid having distractors that could be possible answers (i.e., for a test item such as "where can you find information on Ferguson Jenkins, do not include both "Sports" and "Personalities" as possible answer choices).

3. Incorrect choices will be other page numbers listed in the newspaper index.

4. Some of the incorrect choices should include the same section letter as the correct answer.

Example 14[1]

Applications of Approaches to Norm-Referenced Test Reliability

General Description

Given a description of a situation requiring the interpretation and use of a set of norm-referenced test scores, the student will select from a list of reliability coefficients the coefficient that should be computed, based upon the description of the given situation.

Sample Item

Directions: Read each testing situation described below. Describe which reliability coefficient would be best suited for the situation described. Write the letter preceding the reliability coefficient you have chosen on the separate answer sheet.

Ms. Jones, a ninth grade history teacher, has constructed a final examination in History 9 for the fall semester. She does not have access to either machine scoring facilities or computer analysis. Which reliability coefficient would be best suited for her to compute given the situation?

a. Coefficient of Stability and Equivalence
b. Kuder-Richardson 20
c. Kuder-Richardson 21
d. Coefficient of Equivalence
e. Coefficient of Stability

---

[1] We are grateful to participants at an AERA training program held in Toronto, March 1978, for considerable help in formulating this domain specification.

Stimulus Attributes

1. Each of the test items will consist of 3 parts: (a) a passage describing the norm-referenced testing situation, (b) a question requiring the student to choose which reliability coefficient is best suited, and (c) a set of five possible answers.
2. The passage describing the norm-referenced testing situation will consist of 100 words or less.
   a. The situation described will contain references to paper/pencil tests and performance tests only; no physical diagnostic tests (e.g., hearing) will be described.
   b. The passage describing the testing situation will include:
      i. the purpose of testing
      ii. the area of testing
      iii. whether the test is a group or individual test
      iv. whether the test is standardized, teacher-made, quasi-standardized (i.e., used by all teachers in a school)
   c. The testing situations described may measure either the cognitive, affective, or psychomotor domains.
   d. The situations described are limited to the following categories of use of norm-referenced test scores:
      i. grading
      ii. selection
      iii. placement
      iv. program evaluation
      v. ability grouping
      vi. individual diagnosis

   (See the specification supplement for an expanded discussion of the categories.)

   e. Situations describing the uses of tests in research or in the generation of research hypotheses are not applicable.

BEST COPY AVAILABLE

f. The situations described also involve an explication of the following variables (when appropriate)

   i. speeded or power test
   ii. similarity in difficulty of items (homogeniety of items)
   iii. when test is given (during program or at end)
   iv. the exact nature of the test (aptitude, achievement, or psychomotor)
   v. whether supplemental aids, such as a computer, etc., are available

3. Following the passage describing the norm-referenced testing situation, each item will contain the following question: "Which relaibility coefficient would be best suited for him/her/them to compute, given the situation?"
4. The set of 5 possible answers should be written subject to the restraints set up in the Response Attributes section.
5. All passages should be written at no higher than the 10th grade reading level.

Response Attributes

1. Students will be asked to circle the letter beside one of five possible answers (where one answer is correct and the other four answers are incorrect).
2. The correct answer and the four distractors should be chosen from the following list of reliability coefficients:

   a. coefficient of stability
   b. coefficient of equivalence
   c. coefficient of stability and equivalence

   d. split half (odd-even) reliability coefficient
   e. split half (first half-last half) reliability coefficient
   f. Kuder-Richardson—20 coefficient
   g. Kuder-Richardson—21 coefficient
   h. Inter-rater reliability coefficient

3. For each item, the four distractors chosen will be the four of the seven possibilities that are most nearly suited for the given situation.
4. Answers requiring combinations of reliability coefficients, and "all of the above" and "none of the above" will not be used.
5. The correct answer will be consistent with the discussion and appropriate procedure presented in currently-used educational and psychological measurement texts (i.e., Thorndike and Hagen, Brown, Stanley and Hopkins, Payne, Sax).

Specification Supplement

The following list of situations to be used for developing the test item passages is an expansion of those listed in the Stimulus Attributes. This expansion is necessary to further delimit the content areas.

General Description	Particular Situation
1. Grading	a. classroom grading on a unit of work, no computer-assisted facilities available b. classroom grading on a unit of work, computer facilities available c. end-of-course grading on final exam, no facilities available d. end-of-course grading on final exam, facilities available

BEST COPY AVAILABLE

General Description	Particular Situation
2. Selection	a. selection of a group for a special course, using an achievement test score b. selection of a group for a special activity, using an aptitude test score
3. Placement	a. placement of an individual in an accelerated program, using an achievement test score b. placement of an individual in a special class, based on observation of psychomotor abilities c. placement of an individual in a special group, based on a projective test
4. Program Evaluation	a. evaluation of a program using final test scores
5. Ability Grouping	a. placing students into ability groups based upon i. achievement test scores ii. aptitude test scores iii. performance test scores
6. Individual Diagnosis	a. diagnosis utilizing: i. achievement test score ii. teacher-constructed test score

BEST COPY AVAILABLE

2.5 Item Forms Analysis

Hively, et al. (1973), using the work of Osburn (1968) and Hively, Patterson, and Page (1968) as a basis, have developed a comprehensive method for developing a domain definition called item forms analysis. Based on Osburn's notion of a "universe-defined" test and consonant with Traub's explication of strong domain sampling validity, Hively et al, felt (initially) that their domain definition should satisfy the following two requirements:

1. All the items which could be written from the content domain to be tested must be written (or known) in advance of the final item selection process.
2. A random or stratified sampling procedure must be used in the item selection process.

Before discussing item forms analysis in some detail and also providing some examples of item forms, two comments should be made. One, the experience of Hively et al., in developing item forms, pointed out one very glaring weakness of item generation procedures; they work well only with very structured subject domains, such as mathematics (the subject Hively and his associates considered). Two, and perhaps more important, Hively et al., found that while their attempt to specify "all the behaviors which comprise specific pieces of knowledge" was a great 'quantum leap' over the use of behavioral objectives, it was apparent that it was impossible "to exhaustively define universes of criterion behavior." This forced Hively and his associates into reconsidering the first requirement of their domain, which we listed above. They began to define the sets of test items not as universes of items but as the " nuclei of hypothetical repertoires of behavior," called "domains" (Hively et al, 1973).

Hively and his associates found it an impossible task to list all the items in the content area under consideration, and thus were forced to reconceptualize their approach in terms of domains of behavior, to which a group of items may belong. According to Hively et al.:

> The basic notion underlying domain-referenced achievement testing is that certain important classes of behavior in the reportoires of experts (or amateurs) can be exhaustively defined in terms of structured sets or domainsof test items. Testing systems may be referenced to these domains in the sense that a testing system consists of rules for sampling items from a domain and administering them to an individual (or sample of individuals from a specified population) in order to obtain estimates of the probability that an individual (or group of individuals) could answer any given item from the domain at a specified moment in time.
>
> Domains of test items are structured and built up through the specification of stimulus and response properties which are thought to be important in shaping the behavior of individuals who are in the process of learning to be experts. These properties may be thought of as stratifying large domains into smaller domains or subsets.

Hively et al., use item generation forms to specify these domains of behavior and thus circumvent the problem of trying to exhaustively define the universe on the individual item level. We should note that this switch in conceptualization from "universe of items" to "domain" does not affect the inferential procedures that can be used. If one can develop these domains through the use of item generation forms, the strong domain sampling validity situation (Traub, 1975) will have been attained. The test developer can feel confident in making an inference about what the examinee knows about the domain, based upon his/her test score.

We have discussed the conceptual switch from "universe of items" to "domain" by Hively and his associates for two reasons. One, we are trying to give the reader another sense of the "idealized" nature of being able to specify a content universe and to present a case for how difficult it is to specify a domain. This should reemphasize the practical utility of Popham's domain specification procedure. Two, we have tried to present a context in which to understand the examples that follow.

With these basics specified, we will do the following in terms of item forms analysis. First, we will formally define an item form, as specified by Hively, and discuss three strategies Hively et al., have suggested for developing the domain which the item forms represent. Then, in section 2.6, we will provide an example of an item form and briefly discuss the elements that constitute an item form. This will be done on a cursory level; the reader should refer to the work of Hively et al. (1973) for a detailed discussion. We are here trying to give a "flavor" of the approach; and are not going to do justice to the subtleties. Finally, in section 2.6, we will provide four other relevant examples of item forms, all taken from Hively et al.'s, work on the Minnemast Project.

How does Hively formally define an item form? According to Hively, et al. (1973):

> Items are written as scripts directing the actions of an examiner, with space provided in which to record the responses of a student. Certain elements in the scripts are variable. . .'Item forms' determine the domains of permissable replacements for these variables. By sampling items from these domains, one can estimate the proportion of students who 'have' the system of concepts and skills represented by the item form as a whole, as well as the proportions who respond correctly to various subcomponents.

How does one first develop the general domains in which the item forms serve as item generators? Hively et al (1973), list three possible strategies for developing the domains:

1. Start with a list of prototype items taken from the instructional

material and then alter these items to produce sets of equivalent items measuring the objectives supposedly measured by the prototypical items. Then have content experts review the items so as to end up with a pool of items which purport to measure the instructional objectives.

2. State the instructional objectives and have the item writers develop items which supposedly measure the instructional objectives.
3. Develop hypotheses about sequences and hierarchies of instruction through a careful examination of the basic goals of the instructional unit. Then construct items in accordance with these sequences and hierarchies.

Regardless of the way in which domains are initially defined and developed, at one point, it is necessary that item forms be constructed. Hively _et al_. (1973) give two relevant reasons for the use of item forms:

1. To obviate the necessity to store individual items by substituting a set of written rules through which items can be generated when needed, and
2. to enable the relationships among items to be traced by giving clear specifications of relevant item characteristics.

In other words, the collections of items generated by any of the three procedures just discussed are organized into "formalized schemes," these schemes being the item forms. Each item form is made up of two major parts, one part tells how one would generate the items, the other describes the item's characteristics.

As a means of summing up this discussion of item forms analysis, we can make the following comment. If it is possible to explicitly define the domain, we feel that item generation forms are the mode to use. However, as the reader can see, both from the discussion just presented and the ensuing examples, the complexities of specification can be enormous. Also, the procedures work only for highly structured subject domains. It is for the above two reasons that we prefer the use of Popham's procedure for the development of domain specifications. However, we need to again point out that the domain specification procedure developed by Popham only implicitly defines the domain in question, and the items generated need to be validated by an independent method. That independent method, the use of content specialists, will be discussed in Unit 3.

2.6 Examples of item forms analysis

In this section, we first provide an example of an item form, and then briefly discuss the constituent parts. Then, four more examples of item forms are provided. The first one comes from a paper by Hively *et al* (1968) and the other three are from a book by Hively and his associates (1973)[1].

---

[1] Permission for duplication of these materials in our final report is pending.

Example One

ITEM FORM 2.2*

Producing examples of simple and non-simple, open and closed curves.

GENERAL DESCRIPTION

The child is given an example of a simple open, simple closed, non-simple open, or non-simple closed curve and asked to draw several more that are different, but of the same kind.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

Child is given an example of the required type of curve at the beginning. Child produces curves by drawing them.

Distinguishing Among Cells

Type of curve required: (1) simple open, (2) simple closed, (3) non-simple open, (4) non-simple closed. (The last two curve types are not standard topological classifications, but are clearly defined.)

Varying Within Cells

Instances of sample curves presented.

CELL MATRIX

Script (b)

Simple closed	(1)
Simple open	(2)
Non-simple closed	(3)
Non-simple open	(4)

(Sample curve is drawn from replacement set corresponding to script.)

ITEM FORM SHELL

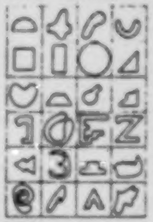
MATERIALS Curve card (a) Response Sheet Pencil	ITEM FORM: 2.2 CELL: 1 REPLICATION: 1
DIRECTIONS TO E Don't look at curve card yourself, until you have laid it in front of S. After S finishes each answer, write its number beside it. If you aren't sure whether S is finished, ask him. In transition to each new question, you can say "um hum" or "O.K." but don't say "good" or otherwise put special emphasis on correct answers.	SCRIPT Here is a (b) simple closed curve. Here is a pencil and paper. Draw another (b) simple closed curve that is different from that one. (Answer #1) Now, draw another (b) simple closed curve that is different. (Answer #2) Now, draw another (b) simple closed curve (Answer #3) Now draw another (b) simple closed curve that is different. (Answer #4)
RECORDING Attach response sheet.	

REPLACEMENT SCHEME

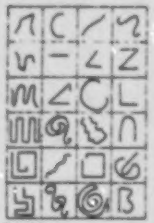
Curve Cards (a)
Cell 1: choose from R.S. 2.1.
Cell 2: choose from R.S. 2.2.
Cell 3: choose from R.S. 2.3.
Cell 4: choose from R.S. 2.4.

Script (b)
Cell 1: simple closed
Cell 2: simple open
Cell 3: non-simple closed
Cell 4: non-simple open

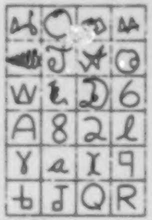
REPLACEMENT SETS



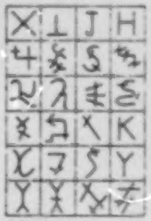
R.S. 2.1.
Simple, closed curves



R.S. 2.2.
Simple, open curves



R.S. 2.3.
Non-simple, closed curves



R.S. 2.4.
Non-simple, open curves

SCORING SPECIFICATIONS

Cell 1 (simple closed): Curve bounds an area, may not have crossing points.

Cell 2 (simple open): Curve does not bound an area, may not have crossing points.

Cell 3 (non-simple closed): Part of curve bounds an area, must have at least one crossing point.

Cell 4 (non-simple open): No part of curve bounds an area, must have at least one crossing point.

---

* Originally developed by Stephen Lundin.

BEST COPY AVAILABLE

What follows is a short description of each of the constituent parts; these descriptions were edited from Hively et al.'s materials.

Item-Form Shell --This element contains the common, unvarying component of all items that could be generated by the item form. The blank spaces in the skill are filled in according to the specifications in the Replacement Scheme. Instructions to the examiner are placed here, and these instructions specify materials, directions, script and recording.

Replacement Scheme --This element specifies how to choose values or prescriptions for each of the variable parts of the item form. Replacements specified in this section come from the Replacement Set.

Stimulus and Response Characteristics --These descriptions are intended to describe and justify whatever behavioral analysis may underlie the properties or characteristics utilized in structuring the domain of items.

Cell Matrix --This element does two things: (1) provides a summary of the information found under Stimulus and Response Characteristics and (2) assigns an identification number to each cell to coincide with the cell numbers used in the replacement scheme.

Scoring Specifications --This section describes the properties to be used to distinguish between correct and incorrect responses.

Now consider the four additional examples of item forms on the next pages.

Example Two

Descriptive Title	Sample Item	General Form	Generation Rules[a]
Basic fact; minuend > 10.	13 — 6	A —B	1. $A=1a$; $B=b$ 2. $(a<b) \epsilon U$ 3. $\{H, V\}$
Simple borrow; one-digit subtrahend.	53 — 7	A —B	1. $A=a_1a_2$; $B=b$ 2. $a_1 \epsilon U - \{1\}$ 3. $(b>a_2) \epsilon U_0$
Borrow across 0.	403 —138	A —B	1. $N \epsilon \{3, 4\}$ 2. $A = a_1a_2 \ldots$; $B = b_1b_2 \ldots$ 3. $(a_1>b_1)$, $(a_3<b_3)$, $(a_4 \geq b_4) \epsilon U_0$ 4. $b_2 \epsilon U_0$ 5. $a_2=0$ 6. $P\{\{1, 2, 3\}, \{4\}\}$
Equation; missing subtrahend.	42—___=25	A—___=B	1. $A=a_1a_2$; $B=b_1b_2$ 2. $a_1 \epsilon U$ 3. $a_2, b_1, b_2 \epsilon U_0$ 4. *Check*: $O<B<A$

[a] Explanation of notation:

Capital letters A, B, . . . represent numerals.

Small letters (with or without subscripts) a, b, a_1, b_2, etc., represent digits.

$x \epsilon \{---\}$: Choose at random a replacement for x from the given set.

$a, b, c, \epsilon \{---\}$: All of a, b, c are chosen from the given set *with replacement*.

N_A: Number of digits in numeral A.

N: Number of digits in each numeral in the problem.

$a_1, a_2, \ldots \epsilon \{---\}$: Generate all the a_i necessary. In general " . . . " means continue the pattern established.

$(a< b) \epsilon \{---\}$: Choose two numbers at random *without* replacement; let *a* be the smaller.

$\{H, V\}$: Choose a horizontal or vertical format.

$P\{A, B, \ldots\}$: Choose a permutation of the elements in the set. (If the set consists of subscripts, permute those subscripted elements.)

Set operations are used as normally defined. Note that $A — B = A — \bar{B}$. Ordered pairs are also used as usual.

Check: If a check is not fulfilled, regenerate all elements involved in the *check* statement (and any elements dependent upon them).

Special sets:

$U = \{1, 2, \ldots, 9\}$
$U_0 = \{0, 1, \ldots, 9\}$

BEST COPY AVAILABLE

Example Three

ITEM FORM 9.7*

Producing a number satisfying a given order relation to specified numbers(s) (spoken form).

GENERAL DESCRIPTION

The child is asked to say the name of a number that bears a specified order relation ("greater than" or "less than") to a given number or numbers in the range 0 through 20. Given numbers are presented in spoken form and response is spoken.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

The presentation is completely spoken; a spoken response is required.

Distinguishing Among Cells

Three scripts are used asking respectively for a number greater than a given number, for a number less than a given number, and for a number greater than one given number and less than another.

Within the third script, three conditions are allowed: (1) first given numeral greater than second with required number possibly an integer; (2) first given numeral greater than second with required number necessarily not an integer; and (3) first given numeral less than second so that the solution to the problem is the empty set.

Varying Within Cells

Within each cell, the given numbers are integers from the range 0 through 20 chosen so that the correct response (when it is not the empty set) can be a real number from the range 0 through 20.

CELL MATRIX

Script (a)	"greater than b_1"	"less than b_1"	"greater than b_1" but less than b_2"		
Numerals (b)	$0 \leq b_1 \leq 19$	$1 \leq b_1 \leq 20$	$0 \leq b_1 \leq 18$ $b_1 + 2 \leq b_2 \leq 20$	$0 \leq b_1 \leq 19$ $b_2 = b_1 + 1$	$1 \leq b_1 \leq 20$ $0 \leq b_2 < b_1$
	(1)	(2)	(3)	(4)	(5)

ITEM FORM SHELL

MATERIALS None	
DIRECTIONS TO E Read script to child. Write down child's exact words.	**SCRIPT** Tell me a number that is ________________

REPLACEMENT SCHEME

(a) Script

Cell 1: "less than b_1" "greater than b_1."
Cells 3,4,5: "greater than b_1 but less than b_2."

(b) Numerals within Script

Cell 1: Choose b_1 from R.S. 9.1
Cell 2: Choose b_1 from R.S. 9.2
Cell 3: Choose two numbers from R.S. 9.3
Cell 3: Choose two numbers from R.S. 9.3
Let b_1 = smaller number; b_2 = larger number
Reject if $b_2 - b_1 \leq 1$

Cell 4: Choose b_1 from R.S. 9.3
Let $b_2 = b_1 + 1$

Cell 5: Choose two numbers from R.S. 9.3
Let b_1 = larger number; b_2 = smaller number
Reject if $b_1 = b_2$

REPLACEMENT SETS

R.S. 9.1: Whole numbers 0,1,2, ...,19.
R.S. 9.2: Whole numbers 1,2,3, ...,20.
R.S. 9.3: Whole numbers 0,1,2. . . . ,20.

SCORING SPECIFICATIONS

Cell 1: Any real number x where $x > b_1$
Cell 2: Any real number x where $x < b_1$
Cell 3: Any real number x where $b_1 < x < b_2$
Cell 4: Any real number x where $b_1 < x < b_2$
Cell 5: Any response equivalent to saying that there are no numbers which can fulfill the conditions.

---

* Originally developed by Donald Sension.

BEST COPY AVAILABLE

Example Four

ITEM FORM 16.14*

Comparing two objects on equal-arm balance and choosing a symbol to complete a statement of the weight relation.

GENERAL DESCRIPTION

The child is asked to compare the weights of two objects that may be (1) indistinguishable by hefting but easily distinguished on the balance, (2) indistinguishable even on the balance. In each of these situations, size varies as an irrelevant dimension. An equal-arm balance is available but instructions for its use are non-directive. The child is asked to select one of the three symbols (>, <, and =) and place it in the blank space provided between the two weight symbols.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

The equal-arm balance is of similar construction to that used in MINNEMAST Unit 16, made of Tinkertoys, cardboard, string, a metal weight, and a foot ruler.

The objects are opaque, cylindrical bottles, identical except for weight (either 23 gm. or 25 gm.) and size (either 2" × ⅝" or 2½" × 1⅜"). Each is identified by a lower-case letter assigned at random.

The child is asked to complete a symbolic statement, corresponding to the weight relation, by choosing the correct relation symbol.

Distinguishing among Cells

Three weight relations (detectable by balance only, not by hefting or "feel") defined in terms of the location of the objects when placed in front of the child:

left > right; left < right; left = right.

Three size relations:

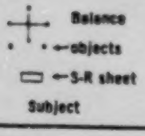
left > right; left < right; left = right.

CELL MATRIX

Weight Relations
(Detectable by Balance Only)

Size Relations	$W_l > W$	$W_l < W_r$	$W_l = W_r$
$S_l > S_r$	(1)	(4)	(7)
$S_l < S_r$	(2)	(5)	(8)
$S_l = S_r$	(3)	(6)	(9)

ITEM FORM SHELL

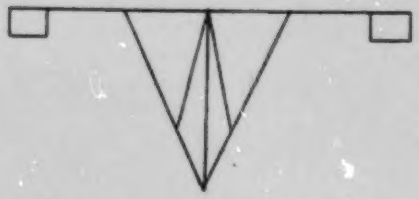
MATERIALS Beam Balance Objects l and r from T.O. 16.14.0 Stimulus-Response sheet (attached) Pencil	
DIRECTIONS TO E Place materials in front of child. (Keep order of objects given above.)  	**SCRIPT** Here are two objects. They have symbols attached to them. Compare them by weight and write one of these three signs (point) in the blank (point) to form the comparison sentence. You may use this balance if you need to.

RECORDING

Attach Stimulus-Response sheet to this page.

Describe what child did.

If balance was used, insert object symbols in schematic drawing of the balance given below, and mark the position of the plumb-line at the time of child's judgment.



DESCRIPTION OF MATERIALS

Pencil (T.O. 16.1.1)

Beam Balance (T.O. 16.13.1): Equal-arm beam balance made from tinker-toy materials as decribed in MINNEMAST Unit 16.

Set of Weight Comparison Objects (T.O. 16.14.0): Set of opaque plastic cylindrical bottles with firmly fitting lids. Two sizes of bottles have been chosen. The small bottle has a length of 2" and a diameter of ⅝". The large bottle has a length of 2½" and a diameter of 1⅜". Two weight values have been chosen so that the objects cannot typically be distinguished by hefting but can be distinguished on the balance. Each object is designated by a randomly chosen, lower-case letter.

Size	Weight 23 gm	25 gm
small	a	m k
large	b	o n

Stimulus-Response sheet (attached to item) (T.O. 16.14.1): a sheet of paper approximately 6" × 4" with the following display:

> Write >, <, or = in the blank
>
> W_l______________W_r

where l and r are the appropriate subscripts (from Replacement Scheme).

REPLACEMENT SCHEME

(l,r) Objects

Cell 1: (o,a)
Cell 2: (m,b)
Cell 3: Choose from R.S. 16.13
Cell 4: (b,m)
Cell 5: (a,o)
Cell 6: Choose from R.S. 16.14
Cell 7: Choose from R.S. 16.15
Cell 8: Choose from R.S. 16.16
Cell 9: Choose from R.S. 16.17

REPLACEMENTS SETS

R.S. 16.13	Ordered pairs	(m,a);	(o,b)
R.S. 16.14	Ordered pairs	(a,m);	(b,o)
R.S. 16.15	Ordered pairs	(b,a);	(o,m)
R.S. 16.16	Ordered pairs	(a,b);	(m,o)
R.S. 16.17	Ordered pairs	(m,k);	(o,n)

SCORING SPECIFICATIONS

A correct response is made by writing the correct symbol (>, <, or =) in the blank space to complete the comparison sentence. This should be > in Cells 1, 2, and 3; < in Cells 4, 5, and 6; = in Cells 7, 8, and 9.

---

* Originally developed by Wells Hively.

BEST COPY AVAILABLE

Example Five

ITEM FORM 26.2*

Plotting a single point on a volume-weight graph.

GENERAL DESCRIPTION

A graph, with axes indicating volume and weight, and a sheet displaying either an ordered pair or a volume-weight chart is presented. The child is asked to plot the point represented by the data onto the grid.

STIMULUS AND RESPONSE CHARACTERISTICS

Constant for All Cells

The grid has the characteristics described in the Description of Materials.

Distinguishing Among Cells

The child is given the data either as an ordered pair or as a Volume-Weight chart.

The data are such that the point to be plotted is either at the intersection of two grid lines, or on an X-axis grid line at a position intermediate (in tenths) between two Y-axis grid lines.

Complete crossing of these categories yields four cells.

Varying Within Cells

The data for the point to be plotted are varied within the limits of the grid and of the Cell Constants specifications.

For Cells 1 and 2, the Volume and Weight values are both chosen from the set of integers 1 through 12, with the requirement that the two values must not be identical. (This condition eliminates situations where order would not matter.)

For Cells 3 and 4, the Volume value is chosen from the set of integers 1 through 12; and the Weight value, $j^{k}/_{10}$ units is chosen so that j is from the set of integers 0 through 11, and k is from the set of integers 1 through 9.

CELL MATRIX

	Y-coordinate an integer	Y-coordinate in tenths
Date as Ordered pair	(1)	(3)
Data as V/W Chart	(2)	(4)

* Originally developed by Graham Maxwell.

ITEM FORM SHELL

MATERIALS Stimulus Sheet (attached) Grid (attached) Pencil	
DIRECTIONS TO E Place materials in front of child and point to the relevant parts as you say: When child has finished, attach both the stimulus sheet and the grid to this page.	**SCRIPT** (d)__________________

DESCRIPTION OF MATERIALS

Stimulus Sheet (attach one of the following objects to the item as specified by (a) in the Replacements).

(T.O. 26.5.1): A sheet of 6"x4" notepaper displaying the ordered pair P [(b), (c)];

(T.O. 26.4.1): A sheet of 6"x4" notepaper displaying the following labeled chart:

OBJECT	VOLUME (in units of volume)	WEIGHT (in units of weight)
P	(b)	(c)

Grid (attached to item) (T.O. 26.2.1): A sheet of paper displaying a grid, 6" × 6", with grid lines ½" apart. On each axis, the grid lines are marked with the numbers 1 through 12. The X-axis is labeled "Volume (in units of volume)," and the Y-axis is labeled "Weight (in units of weight)."

Pencil (T.O. 26.1.1):

REPLACEMENT SCHEME

(a) Stimulus Sheet

Cells 1 and 3:	T.O. 26.5.1
Cells 2 and 4:	T.O. 26.4.1

(b,c) Coordinates of point P for Stimulus Sheet

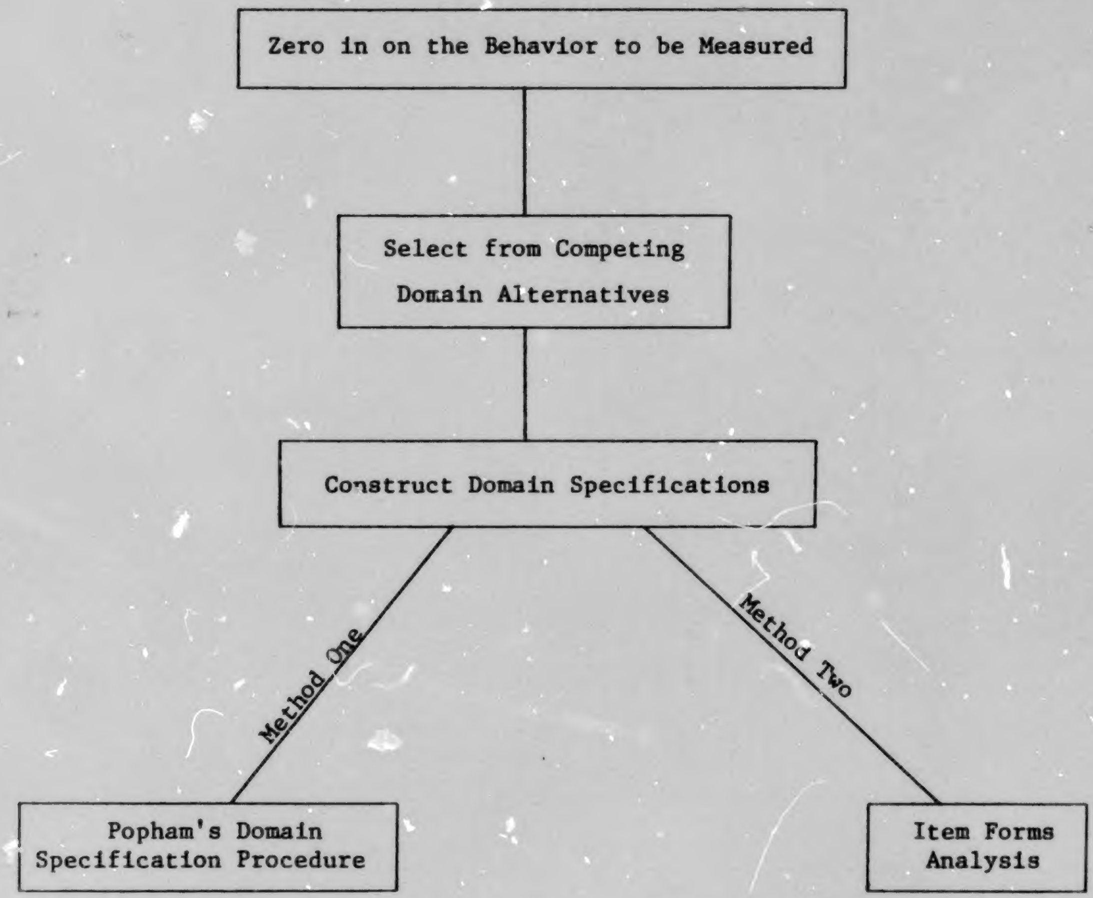
Cells 1 and 2:	Choose b	from R.S. 26.1
	Choose c	from R.S. 26.1
	Reject if b = c	
Cells 3 and 4	Let b = i	
	$c = j^{k}/_{10}$	
	choose i	from R.S. 26.1
	choose j	from R.S. 26.2
	choose k	from R.S. 26.3

BEST COPY AVAILABLE

2.7 Flowchart of the Process

Figure 2.7.1 should be helpful since it provides a summary of the steps a criterion-referenced test developer must consider in preparing domain specifications.

Figure 2.7.1. Steps for preparing domain specifications.


Zero in on the Behavior to be Measured
Select from Competing Domain Alternatives
Construct Domain Specifications
Method One
Method Two
Popham's Domain Specification Procedure
Item Forms Analysis

2.8 Objective Banks

There are presently available a number of commercially prepared sets of objectives that can be utilized by a criterion-referenced test developer in his/her work. The test developer may find, however, that he/she must work further with these objectives as they are usually stated in behavioral terms, and lack sufficient clarity to permit a clear determination of the domain of test items intended for the objectives (see section 2.2). These sets of objectives may serve as an excellent starting point in the development of domain specifications.

The addresses of several of the better known organizations that distribute objectives sets are given below. These organizations provide complete catalogs of subject areas for which objectives have been prepared, and certain of the organizations provide listings of supplemental services that can be used in conjunction with the objectives sets.

Instructional Objectives Exchange (IOX)
Box 24095
Los Angeles, California 90024

Westinghouse Learning Press Publications
770 Lucerne Drive
P. O. Box 9035
Sunnyvale, California 94086

Other organizations distributing objectives (and/or test items) can be located in the classified ads section of Phi Delta Kappan.

2.9 Preparation of Test Specifications

Any good test requires planning. In this section, we will discuss a series of steps that will aid in the planning process. These steps are involved with the preparation of a set of test specifications. In turn, the test specification stage may be viewed as the initial step in the development of a criterion-referenced test. Thus, we will be discussing a series of steps that will aid in the subsequent development of a criterion-referenced test. According to Tinkelman (1971), whose work we have utilized extensively in preparing this section:

> The essence of initial test planning is establishing the test specifications; that is, the sum total of the qualities and characteristics that the test should possess.

The following list of steps to guide in the development of test specifications was taken from Tinkelman (1971) and adapted to fit a discussion of criterion-referenced tests. We present the steps first, and then comment on them. The reader will find that certain of the steps have been covered in detail in other sections.

Steps in Developing Test Specifications

1. Define the general purpose and requirements of the test.
2. Establish the specific scope of the test as expressed by the domain specifications or item forms.
3. Select appropriate item types.
4. Determine the appropriate number of test items to be used.
5. Establish how items are to be assembled in the test.
6. Prepare item-writing and item-review assignments.

Steps 2 and 3 were discussed in detail in earlier sections of the unit. The other steps will be given greater emphasis in the ensuing discussion.

Step 1: Define the general purpose and requirements of the test.

According to Tinkelman (1971), the test developer should try to answer the following questions in order to clarify his/her general purpose for testing:

1. What specific content areas are to be measured?
2. Who is to be tested?
3. How are the test scores to be used?
4. What are the time limitations on testing?
5. Will there be a need for equivalent forms?

A number of other possible questions can be asked; the point of the process is to get the test developer to zero in on what the purpose and requirements of his/her test are going to be.

Step 2: Establish the specific scope of the test as expressed by the domain specifications or item forms.

A great deal has been written in this unit on the development of domain specifications (sections 2.3 and 2.4) and/or item forms (sections 2.5 and 2.6). The only point to be made here is that, in the overall context of the development of a set of test specifications, the domain specification phase is the second step in the process.

Step 3: Select appropriate item types.

This step has been discussed in detail in section 2.10 and also in section 4.3. According to Tinkelman, the item type or item types to be chosen should be considered in reference to:

1. the domain specifications or item forms
2. possible scoring procedures
3. administrative features
4. printing requirements.

The list is in order of priority; however, a consideration of all four may be necesary before making a final decision on which item type or types to use. It should be pointed out again that first and foremost, the item type chosen must be such that the items do indeed "tap" the behavior specified in the domain definition. All other considerations are secondary.

Step 4: Determine the appropriate number of items to be used.

At this point in the planning process, the test developer is trying to get an indication of the number of test items that will be needed. This then will have a bearing on the number of items that item writers need to construct. Four areas should be considered in making tentative decisions about number of items:

1. The relationship of numbers of items to the importance placed upon the domain in the curriculum.
2. The relationship of the numbers of items to minimum reliability requirements.
3. The relationship of the number of items to time limits.
4. The relationship of the number of items to item-review mortality rate.

In terms of area number one, it may be the case that certain areas of a curriculum have been stressed more than others in the instructional process. If the test developer plans for the test to cover multiple domains, he/she should then plan, when drawing samples of items from each domain, to more heavily sample the most important domains. Such a decision is situation specific, and little more can be said in terms of overall guidelines.

In reference to area two, the relationship of the number of items to minimum reliability requirements, guidelines are presently being developed. As discussed in the section of unit 5 on reliability, the Spearman-Brown formula, which relates test length to reliability, is reasonable to use only for norm-referenced tests. Similar relationships need to be developed for two important uses of criterion-referenced test scores, domain score estimation and assignment of examinees to mastery states. The following procedure should be helpful to those in the planning process for determining test length when domain score estimation is the problem of interest. The solution is a conservative one, i.e., test lengths determined by this method will be a little longer than they need to be to obtain the degree of precision required by the test developer. The formula[1] is:

$$\boxed{\text{Test Length} = \frac{.25}{(\text{degree of precision})^2}}$$

Ask yourself (or interested others): What degree of precision is required of the domain score estimates? Discuss the degree of precision question in the same way you would the standard error of measurement. A primary difference between the two is that domain score estimates are defined on a scale [0, 1].

---

[1] The formula can be derived from the binomial test model.

Example

Suppose you felt that an error of ± 10% could be tolerated; then, degree of precision = .10; and, using the equation above, test length = 25.

There is one other important consideration: Item review mortality rate. You must try and estimate the percentage of items that are likely to be discarded in the review process. Ask yourself: How experienced are my item writers? A rejection rate in the neighborhood of 20% would not be unusual. This figure may seem especially low. It certainly is by norm-referenced test development standards, but, the "standards" for a good criterion-referenced test item, while being difficult to meet, do not depend on desirable "statistical properties", something which is important for norm-referenced test items. This is something that is very hard to predict in advance by norm-referenced item writers. Hence, we have a good explanation for the relatively higher rejection rate of items prepared for norm-referenced tests than criterion-referenced tests.

Continuing the example, if in the judgment of the test developer, about 20% of the item pool for an objective is apt to be poor, then we must write about 31 test items. (Solution: Let the number of test items prepared be X. If X - (20% of X) = 25; then, X $\doteq$ 31.) Item writers would need to prepare about 31 test items.

Two points seem worthy of mention at this point, One, it is unlikely that fewer than five or six items measuring an objective will produce desired levels of reliability. Two, while no tables or formulas exist to connect test length to reliability (or consistency) of decision-making, this can be studied empirically after the administration of a pool of test items. "Post-hoc" test forms of varying lengths can be constructed and reliability estimates may be calculated, on the assumption that examinees would have responded in the same way had they been presented with the "parallel-forms" rather than a single large pool of test items. By varying the length of the forms and the formation of parallel-forms (i.e., which items are placed in which forms), the relationship between test length and reliability for a specified sample of examinees for a pool of test items measuring a particular domain specification can be studied.

In reference to area three, it really goes without saying that the number of items needed is determined by time limits. However, it should be noted that this also depends on the item type or types chosen in step 3. For instance, more true-false questions can be asked in a particular time period than completion questions. We can offer few general guidelines; the decision will depend upon the content area, the students tested, the item type(s) selected, and the total time available.

Finally, the number of items needed is dependent on the item-review mortality rate, that is, the number of items that can be expected to be rejected either because of technical flaws or content validity problems. Clearly, the determination of a figure will be situation-specific.

Step 5: Establish how the items are to be
assembled in the test.

The material in section 4.5 is relevant in considering this step of the test specification procedure. Therefore, the reader should refer to section 4.5 for details.

Step 6: Prepare the item-writing and item-review assignments

The item-writing assignments can be handled in three general ways:

1. An item-writer can concentrate on developing items for a single or a few domain specifications; or
2. if the writer has an item type specialty, he/she can concentrate on the same item types across domains; or
3. if item-writer "staleness" becomes a problem, the item-writer can work on a number of domains.

Of course, many offshoots of these very general guidelines are possible. Further, if the test developer is also the item writer, which is often the case, then the guidelines above are of little use.

2.10 Preparation of Test Items

In preparing criterion-referenced test items, one important point should be kept in mind: Close attention must be given to the domain specifications at all times, thereby insuring that the test items "tap" behaviors in the domain of behaviors defined by the specification. With this in mind, the materials included in this section should help the reader to choose the relevant type of test item to suit his/her purpose, and then to make sure that proper item writing principles are followed in item preparation.

Fortunately, a considerable amount of literature exists to (1) introduce practitioners to available item formats, (2) help practitioners select the "best" item formats to measure particular objectives, and (3) train practitioners to write "good" test items ("good" in the sense of being technically correct and measuring the intended objective). The interested reader is referred to section 2.12.3 for an excellent selection of references. In this section, several summaries are included:

1. Types of item formats (a list of item formats for objectively and subjectively scored test items),
2. Definitions and appropriate item formats for objectives classified into different levels of Bloom's Taxonomy of Educational Objectives, Cognitive Domain,
3. Some differences between essay and objective tests (a comparison of these two common types of tests on nine dimensions),
4. Principles of item writing (a list of questions, organized by item format, concerning the quality of test items),
5. Scoring of objective and essay test items.

Item Formats

- Objectively Scored Items
  - Multiple-Choice
  - Matching
  - True-False
  - Short Answer/Completion
- Subjectively Scored Items
  - Essay
  - Performance

Definitions and Appropriate Item Types for Objectives Classified into Different Levels of Bloom's Taxonomy of Educational Objectives, Cognitive Domain

Category	Definition of the Ability Involved	Verbs Typically Used to Describe Objectives	Possible Test Item Formats
Knowledge	Knowledge of specifics, terminology, specific facts, ways and means of dealing with specifics, conventions, trends, sequences, classifications and categories, criteria, methodology, universals and abstractions in a field, principles and generalizations, and theories and structures.	define describe identify recall recognize name state recite write acquire label list	Multiple-Choice Matching True-False
Comprehension	A type of understanding such that the person knows what is in a message and can use the information without connecting it necessarily to other pieces of information or understanding the fullest implications of the message.	translate transform give in own words illustrate prepare rephrase restate represent explain interpret	Multiple-Choice Matching True-False
Application	Involves the use of abstractions (e.g., rules or ideas) in concrete situations.	apply generalize relate develop organize use transfer demonstrate compute solve produce employ	Multiple-Choice Matching

Category	Definition of the Ability Involved	Verbs Typically Used to Describe Objectives	Possible Test Item Formats
Analysis	Decoding communication into the proper elements so as to reveal their relationships.	distinguish classify discriminate analyze contrast deduce subdivide identify differentiate compile categorize create summarize arrange	Short Answer Completion Essay
Synthesis	Placing elements together to form a whole, when the whole was not clear before.	write tell modify specify produce combine synthesize categorize create organize	Short Answer Completion Essay
Evaluation	Determining the worth of some material for a given purpose or use.	judge assess decide compare contrast standardize appraise criticize conclude interpret	Completion Essay

Some Differences Between Essay and Objective Tests

Essay	Objective
1. Student plans his/her own answer and expresses his/her own beliefs.	1. Student selects an answer or provides a short answer.
2. The test includes relatively few, usually general questions, calling for extended answers. It covers less of the curriculum, but the part covered is in-depth.	2. The test contains many specific questions requiring brief answers. It covers more of the curriculum, but it is in less-depth.
3. Thinking and writing time is needed.	3. Thinking and reading time is needed.
4. The quality of the test is determined mainly by the skill of the person grading the paper.	4. The quality of the test is determined mainly by the test constructor.
5. The test is relatively easy to prepare, but tedious and difficult to score.	5. The test is tedious and difficult to prepare, but easy to score.
6. Much freedom is given to the student to express his/her ideas in his/her own words.	6. The test constructor has freedom to express his/her own values and preferences. The student has freedom to show, by his/her score, knowledge of test content.
7. The student can bluff.	7. The student can guess.
8. It is less clear to the student what is expected in an answer.	8. It is quite clear what is expected of the student.
9. The distribution of test scores is determined by the person grading the papers.	9. The distribution of test scores is determined by the test constructor.

---

[1] From Hambleton, R. K., and Fitzpatrick, A. Review techniques for criterion-referenced test items. (In preparation)

Objectively-Scored Item Writing Principles[1]

General Principles

1. Assess only a single piece of knowledge or skill in a test item.

2. Test item readability should be at level appropriate for the examinees being tested.

3. Avoid the use of "trick" test items or test items measuring minor or insignificant points.

4. Always identify the source of opinions or quotes used in test items.

5. Avoid measuring knowledge or skills in a test item which are extraneous to those which the test item was written to measure.

6. Remove superfluous words or complications in a test item which will introduce irrelevant factors into examinee test performance.

7. Test items must be written clearly.

8. Test items must be constructed in accord with standard rules of punctuation and grammar.

9. Negatives should be underlined or highlighted in some way.

10. Avoid the use of words which give clues to correct answers.

11. A test item must have one correct or clearly best answer.

12. Examinees who have the skill or knowledge measured by a test item must answer it correctly.

13. Insure that the correct answers follow a random pattern.

14. Have content and measurement specialists review test items to eliminate ambiguity, technical errors, and other item writing errors.

---

[1] We would like to thank Anne Fitzpatrick for assistance in this section of the Unit.

Writing Multiple-Choice Test Items

	Yes	No	Unsure
Item Stem Content:			
1. Has new material been used in the item if it measures students' understanding or their ability to apply principles?	✓		
2. Could the item be better expressed as a series of true-false questions?		✓	
3. Is the content of the test item reflective of the domain specification the item was prepared to measure?	✓		
4. Does the item stem clearly define a problem?	✓		
Item Stem Structure:			
1. In the item, is as much material as possible included in the stem so that the options are as short as possible?	✓		
2. Have all repetitive words or phrases been placed in the item stem rather than in the set of answer choices?	✓		
Response Content:			
1. Is there only one correct or one best answer to each item?	✓		
2. If the best answer form is used, are the distractors clearly less correct than the "best" answer?	✓		
3. If the correct answer form is used, are the distractors of an item clearly incorrect?	✓		
4. Will all the distractors to the item be plausible to those who do not possess the skill measured by the item?	✓		
5. Does the set of answer choices for the item contain a vocabulary or reading load which will act as irrelevant sources of difficulty?		✓	

	Yes	No	Unsure
Response Structure:			
1. Is the number of distractors for the item appropriate to the ages of those being tested?	✓		
2. Have possible answers such as "all of the above" and "none of the above" been avoided in the item?	✓		
3. Does the item contain two or more distractors which overlap or mean the same thing, such that an examinee could eliminate these distractors simultaneously?		✓	
4. Are all possible answers to an item similar in type, concept or focus so that they are as homogeneous as possible?	✓		
5. Are all possible answers of the item grammatically consistent with the item stem?	✓		
6. Do the answer choices of the item have the same grammatical form so that they are parallel?	✓		
7. Are the possible answers to the item similar in length and complexity?	✓		
8. Are the possible answers to the item listed on separate lines below the item stem?	✓		
9. Are the possible answers to the item arranged in a logical order where possible?	✓		
10. Are letters used in front of the possible answers to identify them?	✓		
11. Has "don't know" been used as an answer choice?	✓		
Directions:			
1. Do directions to the test clearly specify whether the correct or whether the best answer is to be chosen?	✓		

Writing Matching Test Items

Item Stem (Premise) Content:

	Yes	No	Unsure
1. Do the matches to be made in an item all reflect important aspects of the subject material to be tested?	✓		
2. Are all premises clear in meaning?	✓		
3. Are the premises of the item clearly related to one another so that they are as homogeneous as possible?	✓		

Item Stem (Premise) Structure:

	Yes	No	Unsure
1. Are all the premises of a set similar in grammatical form?	✓		
2. Does any set of premises have more than 8-10 elements?		✓	

Response Content:

	Yes	No	Unsure
1. Do any of the premises plausibly relate to a response other than its correct match?		✓	
2. Are the responses to the item similar in type, focus or concept, so that they are as homogeneous as possible?	✓		
3. Can correct matches be made by using only logic or a superficial understanding of the subject material?		✓	

Response Structure:

	Yes	No	Unsure
1. Are there more responses than premises?	✓		
2. Are the responses arranged in a systematic order wherever possible?	✓		

	Yes	No	Unsure
3. Do short phrases, words or numbers make up the response list of an item whenever possible?	✓		
4. Do the responses share the same grammatical form?	✓		
Directions:			
1. Are there headings for the premise and response lists of an item?	✓		
2. Do the directions clearly specify the basis on which matches are to be made?	✓		
3. Do the directions clearly state whether a response can be used more than once?	✓		
4. Is the matching exercise presented on a single page?	✓		

Writing True-False Test Items

	Yes	No	Unsure
Item Content:			
1. Can it be said without qualification that the item is definitely true or false?	✓		
2. Is only a single idea expressed in the statement comprising the item?	✓		
3. Is any part of the item true, while another part of that item is false?		✓	
Item Structure:			
1. Is the item statement short?	✓		
2. Is the sentence structure of the item statement simple?	✓		
3. Is each item stated as concisely as possible?	✓		
4. Does the item contain vague words like "seldom," "frequently," or "generally"?		✓	
Response Content:			
1. Could a person use simply logic or common sense to identify the correct answer?		✓	
2. Will the wrong answer to an item be plausible to those who have not mastered the subject material?	✓		
Directions:			
1. Are directions included which clearly describe how examinees should answer the items?	✓		

Writing Short Answer/Completion Test Items

	Yes	No	Unsure
Item Content:			
1. Does the item pose an important rather than a trivial question about the subject matter?	✓		
2. Is the item written so clearly that there is a single correct answer which a good student will know?	✓		
3. Is the item written so that a brief answer is possible?	✓		
4. Is the meaning of the item made unclear because of too many blanks in the item?		✓	
Item Structure:			
1. Have response cues or specific determiners such as "a" and "an" or singular and plural verbs been avoided in the item?	✓		
Response Content:			
1. Is the student asked to provide only key words, phrases or sentences in response to the item?	✓		
2. Is the precision desired in the answer to the item clearly indicated?	✓		
3. If the item requires a numerical answer, are the units of the answer specified?	✓		

	Yes	No	Unsure
Response Structure:			
1. Are the blanks which the student will fill in placed near the end of the item?	✓		
2. Has ample space to record an answer been provided in the item?	✓		
3. Are the answer spaces provided for the items all the same length?	✓		
Directions:			
1. Is it clearly indicated what form the answers to the items should take?	✓		
2. Is it clearly stated whether spelling and grammatical errors will be scored?	✓		
3. Are students informed of how the answers will be scored?	✓		

Writing Essay Test Items

	Yes	No	Unsure
Item Content:			
1. Does the item pose a clear task for the examinee by including a clear specification of the scope and direction desired in an answer?	✓		
2. Are students asked, in the item, to "compare," "contrast," "give the reason for," "explain how," etc., rather than simply state "what," "where," "when," "who," or "where"?	✓		
3. Does the essay question pose a new problem to the examinees?	✓		
4. Is the item a general or broad question which could be better expressed by several more concise questions?		✓	
Item Structure:			
1. Is the question of unsuitable length or complexity for the maturity levels of the students?		✓	
Directions:			
1. Are students informed of an appropriate amount of time they should spend on each essay?	✓		
2. Are students informed of the number of points associated with each essay question?	✓		
3. Are students informed of how their responses to the items will be scored?	✓		
4. If students are permitted to choose which of several questions to answer, are these several questions equal in difficulty?	✓		
5. Has an "ideal" response to each question been prepared before test administration?	✓		

Item Scoring

Multiple Choice Test Items

The most commonly used formula for correcting for guessing is:

$$CS = R - W/\ (n-1)$$

Where CS is the score corrected for guessing
R is the number of correct answers
W is the number of incorrect answers
(not counting omitted questions)
n is the number of choices for each item.

Examples

1. - for two choice questions, n = 2

$$CS = R - W/\ (2-1) = R-W.$$

- for a 90 question test in which student had 60 correct answers, 10 wrong and 20 omits - $CS = 60 - 10 = 50.$

2. Usually we have 5 choice questions, n = 5

then $CS = R - W/\ (5-1) = R - W/4$

- for a 90 question test in which student had 50 correct answers, 30 incorrect and 10 omits

$$CS = 50 - (30/4) = 42.5$$

To see how this formula corrects for guessing - suppose a student takes a 5 choice-90 question test. If the student were to guess at each question - for each question he would have one chance in five of obtaining the correct answer. Thus on 90 questions, he would obtain 18 correct answers by guessing (1/5 x 90 = 18). Hence he would obtain 72 incorrect answers (90 - 18 = 72). Applying the correction - for - guessing formula:

$$\begin{aligned} CS &= R - W/4 \\ &= 18 - (72/4) = 0 \end{aligned}$$

Now suppose a student is able to answer 50 of the questions on the basis of knowledge and to the remaining 40 questions he guesses randomly. He would obtain 50 correct answers from knowledge and 8 more by guessing. (If he guesses to 40 questions, on the average he should get 8 right by chance - (1/5 x 40 = 8). Thus he has 58 correct answers and 32 incorrect answers giving him a corrected score of:

$$\begin{aligned} CS &= 58 - 32/4 \\ &= 58 - 8 \\ &= 50 \end{aligned}$$

It is clear then that guessing at random will not improve your score when the correction - for - guessing formula is applied.

Two suggestions are proposed:

1. Students should be informed of the correction - for - guessing formula to be used so they can formulate a strategy for writing the test.

2. Use of the correction - for - guessing formula is very important when the test is speeded (when most people do not finish the test) because it eliminates a large amount of wild guessing. It is relatively ineffective when most students have time to answer all of the test questions.

At this point there seems to be little evidence to recommend "correction-for-guessing" scoring with criterion-referenced tests. Generally, there seems to be less guessing on criterion-referenced testing because of the instructional relevancy of the tests. Also, since the emphasis in test score interpretation is not on a comparison of students, there is less pressure on students to achieve high scores. Finally, when instructional decisions are to be made, examinees far from a cut-off score will be unaffected by a correction-for-guessing formula. For those examinees near the cut-off score (which is usually in the region of 70% to 90%), the amount of guessing will be minimal. Therefore, there seems to be little value for applying a "correction-for-guessing." We do see merit however in two suggestions:

1. The "don't know" answer should be considered as an an answer choice to reduce the effects of guessing,

2. Adjust cut-off scores upward to reduce the chance that examinees will "demonstrate" mastery because they were lucky enough to guess the answers to a few questions.

Scoring

Short Answer/Completion Items

	Yes	No	Unsure
1. Has enough time been set aside for scoring the tests?	✓		
2. For each item, has the answer or set of answers which should be considered correct been identified?	✓		
3. Have variations of the correct answer, which might be considered partially correct, been identified?	✓		
4. Has the manner in which correct answers will be scored been identified?	✓		
5. Has the manner in which partial credit will be given for a response been identified?	✓		
6. Has a scoring system for each aspect of a response such as spelling or grammar been specified, if these qualities are to be assessed?	✓		
7. Has a scoring key been prepared, if needed?	✓		
8. Will the scoring key be checked against a random sample of completed tests to make sure that the key accommodates all interpretations of each question?	✓		
9. Will people who have mastery in the subject area of the test be scoring the tests?	✓		
10. Will a complete set of correct answers be provided to each student who takes the test?	✓		

Essay Test Items

	Yes	No	Unsure
1. Have arrangements been made with two, or preferably more, readers to independently evaluate each of the essays?	✓		
2. Will the readers selected be skilled in the content areas to which the essay questions relate?	✓		
3. Will each reader be asked to evaluate all reponses to one question at a time?	✓		
4. Will essay readers be given enough time to grade all responses to a question without interruption?	✓		
5. Will all answers to a question be as anonymous as possible?	✓		

	Yes	No	Unsure
6. Will readers be advised to shuffle all answers to a question before beginning their evaluations?	✓		
7. Has a uniform grading system been established which will apply to all responses to a question?	✓		
a. Analytical method:			
1. Has an "ideal" answer to each essay question been prepared?	✓		
2. Have the contents of each ideal answer been identified and listed?	✓		
3. Have other qualities of each "ideal" response such as logical organization, grammar, support of statements, etc. been identified and listed?	✓		
4. Has each aspect of content and other qualities been assigned score points?	✓		
5. Has a procedure been established to indicate partial demonstration of a listed aspect?	✓		
6. Will each reader record, for each essay, the presence or absence of each listed aspect contained in that essay?	✓		
7. Will a person other than one of the readers be responsible for assigning score points to the essays evaluated by the readers?	✓		
b. Global method:			
1. Have the categories into which an essay is to be classified, in terms of its overall quality, been specified?	✓		
2. Have actual (or sample responses) to represent each of the several categories been identified (or devised)?	✓		
3. Will all essay readers read, rate and discuss each essay, representing a category?	✓		
4. Will each essay be read rapidly and a global impression of its quality be used to classify it?	✓		
5. Will at least two readers read and classify each essay in terms of its overall quality?	✓		
6. Will the sum or the average of the ratings of an essay be used as a final score for that essay?	✓		

2.11 Editing Test Items

At this point in the development process, it is important for the test constructor to check the test items developed to see if they meet the basic technical criteria set for items. The focus at this point should be on the technical quality of the items and the suitability of the directions to the student about how to respond. At a later time point, other reviewers will be asked to comment on the content validity of the items.

The item review form presented on the next three pages will be helpful to individuals interested in conducting a systematic technical analysis of their items. Two points are worthy of mention. One, our item review form is specific to multiple-choice test items. Of course, it will be quite easy for anyone to use our format and principles for preparing other types of items, and design new item review forms, one for each item type. Two, section two of the item review form was designed to be content specific. In this instance, the area was reading/language arts. In different content areas, it is likely that other relevant questions would be included in section two. Other times, section two may be deleted. The item review form was used recently (in a modified form) at an item writing workshop in the Montgomery County Public School System (MCPS) (Rockville, Maryland). The workshop was conducted by the two authors with the excellent assistance of Lois Martin, Kay Morgan, and Liz Flach from MCPS. We are grateful to them (and many of their colleagues) for their constructive criticisms and helpful comments.

Item Review Form (Multiple-Choice)

Objective Number: ______________ Test Item Number: ______________

Reviewer: ____

Date: ________________________

- -

Objective:

Test Item:

- -

Section I. Technical Quality

Place a "✓" under the column corresponding to your rating of the test item for the questions in this section and the next one.

	Yes	Questionable	No
1. Is the item stem clearly written for the intended group of examinees?	___	___	___
2. Is the item stem free of irrelevant material?	___	___	___
3. Is a problem clearly defined in the item stem?	___	___	___
4. Are the choices clearly written for the intended group of examinees?	___	___	___
5. Are the choices free of irrelevant material?	___	___	___

	Yes	Questionable	No
6. Is there a correct answer or a clearly best answer?	___	___	___
7. Have words like "always," "none," or "all" been removed?	___	___	___
8. Are likely examinee mistakes used to prepare incorrect answers?	___	___	___
9. Is "all of the above" avoided as a choice?	___	___	___
10. Are the choices arranged in a logical sequence (if one exists)?	___	___	___
11. Was the correct answer randomly positioned among the available choices?	___	___	___
12. Are all repetitious words or expressions removed from the choices and included in the item stem?	___	___	___
13. Are all of the choices of approximately the same length?	___	___	___
14. Do the item stem and choices follow standard rules of punctuation and grammar?	___	___	___
15. Are all negatives underlined?	___	___	___
16. Are grammatical cues between the item stem and the choices, which might give the correct answer away, removed?	___	___	___
17. Is the item format appropriate for measuring the intended objective?	___	___	___
18. Does the test item measure the intended objective?	___	___	___
19. Does the test item measure only the intended objective?	___	___	___

Section II. Technical Quality Matters Specific to Reading/Language Arts Test Items

	Yes	Questionable	No
1. Can a correct answer be given without reading the passage?	___	___	___
2. Is the discourse appropriate for measuring the intended objective?	___	___	___

	Yes	Questionable	No
3. Does the discourse and test item provide a valid measure of the intended objective?	___	___	___
4. Do the following fall within the range for the number of words in each sentence of the			
(a) directions?	___	___	___
(b) discourse?	___	___	___
(c) item stem?	___	___	___
(d) item choices?	___	___	___
5. Do the following fall within the range for the number of sentences in the			
(a) directions?	___	___	___
(b) discourse?	___	___	___
(c) item stem?	___	___	___
(d) item choices?	___	___	___
6. Is there the desired number of words in the selection of discourse?	___	___	___
7. Does the test item contain the desired number of choices?	___	___	___
8. Is the ratio of common to uncommon words correct?	___	___	___

- -

Suggested Revisions:

- -

Final Rating (Check One):

☐ Accept

☐ Accept (with revisions- see above)

☐ Reject

2.12 References

The references are divided into three sections: References Cited, References for Further Study, and Measurement and Evaluation Textbooks.

2.12.1 References Cited

Allendoerfer, C. B. The utility of behavioral objectives: A valuable aid to teaching. *Mathematics Teacher*, December 1971, *686*, 738-742.

Anderson, R. C. How to construct achievement tests to assess comprehension. *Review of Educational Research*, 1972, *42*, 145-170.

Berk, R. A. The application of structural facet theory to achievement test construction. *Educational Research Quarterly*, 1978, *3*, in press.

Bormuth, J. R. On the theory of achievement test items. Chicago: University of Chicago Press, 1970.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington: American Council on Education, 1971.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: John Wiley & Sons, 1972.

Duchastel, P. E., & Merrill, P. F. The effects of behavioral objectives on learning: A review of empirical studies. *Review of Educational Research*, 1973, *43*, 53-69.

Ebel, R. L. Evaluation and educational objectives. *Journal of Educational Measurement*, 1973, *10*, 273-279.

Forbes, J. E. The utility of behavioral objectives: A source of dangers and difficulties. *Mathematics Teacher*, December 1971, *687*, 744-747.

Gagne, R. M. Behavioral objectives? Yes! *Educational Leadership*, February 1972, 394-396.

Hambleton, R. K. Applications of latent trait models to the development and uses of criterion-referenced tests. *Laboratory of Psychometric and Evaluative Research Report No. 91*. Amherst, MA: School of Education, University of Massachusetts, 1979.

Hively, E., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project. CSE monograph series in evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Hively, W., Patterson, H. L., & Page, S. A. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-290.

Kneller, G. F. Behavioral objectives? No! Educational Leadership, February 1972, 397-400.

MacDonald, J. B., & Wolfson, B. J. A case against behavioral objectives. The Elementary School Journal, December 1970, 119-128.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Co., 1974.

Millman, J. Hang the hang-ups about test making. A paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, "Criterion-Referenced Measurement: The State of the Art," Washington, D.C., October 27, 1978.

Osburn, H. G. Item sampling for achievement testing. Educational and Psychological Measurement, 1968, 28, 95-104.

Popham, W. J. Probing the validity of arguments against behavioral goals. A symposium presentation at AERA, Chicago, Illinois, 1968.

Popham, W. J. An approaching peril: Cloud-referenced tests. Phi Delta Kappan, 1974, 56, 614-615.

Popham, W. J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1978. (a)

Popham, W. J. A lasso for runaway test items. A paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, "Criterion-Referenced Measurement: The State of the Art," Washington, D.C., October, 1978. (b)

Scandura, J. M. Problem-solving: A structural/process approach with educational implications. New York: Academic Press, 1977.

Tinkelman, S. N. Planning the objective test. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Traub, R. E. Stirring muddy water: Another perspective on criterion-referenced measurement. Ontario Institute for Studies in Education, mimeo, 1975.

2.12.2 References for Further Study

Baker, E. L. Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. Educational Technology, 1974, 14, 10-16.

Coffman, W. E. Essay examinations. In Thorndike, R. L. (Ed.) Educational Measurement. (2nd edition) Washington, D.C.: American Council on Education, 1971.

Jackson, R. Developing criterion-referenced tests. TM Report No. 1. Princeton, N.J.: ERIC Clearing House on Tests, Measurement and Evaluation, 1970.

Nitko, A. J. Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California,.1974.

Popham, W. J. (Ed.), Criterion-referenced measurement: An introduction. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Posner, G. J., & Strike, K. A. A categorization scheme for principles of sequencing content. Review of Educational Research, 1976, 46, 665-690.

Roid, G. H., & Haladyna, T. M. A comparison of objective-based and modified-Bormuth item writing techniques. Educational and Psychological Measurement, 1978, 38, 19-28.

Shoemaker, D. M. Toward a framework for achievement testing. Review of Educational Research, 1975, 45, 127-147.

Wesman, A. G. Writing the test item. In Thorndike, R. L. (Ed.) Educational Measurement. (2nd edition) Washington, D.C.: American Council on Education, 1971.

2.12.3 Measurement and Evaluation Textbooks

Ahmann, J.S., and Glock, M.D. *Evaluating pupil growth*. (5th ed.) Boston: Allyn and Bacon, 1975.

Anastasi, A. *Psychological testing*. (4th ed.) New York: Macmillan, 1976.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill, 1971.

Chase, C.I. *Measurement for educational evaluation*. Reading, MA: Addison-Wesley, 1974.

Cronbach, L.J. *Essentials of psychological testing*. (3rd ed.) New York: Harper and Row, 1970.

Ebel, R.L. *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Gronlund, N.E. *Measurement and evaluation in teaching*. (3rd ed.) New York: Macmillan, 1976.

Hills, J.R. *Measurement and evaluation in the classroom*. Columbus, OH: Charles E. Merrill, 1976.

Lemke, E., and Wiersma, W. *Principles of psychological measurement*. Chicago: Rand McNally, 1976.

Lewis, D.G. *Assessment in education*. New York: Wiley, 1975.

Lien, A.J. *Measurement and evaluation of learning*. (3rd ed.) Dubuque, IA: Wm. C. Brown Company, 1976.

Lindvall, C.M., and Nitko, A.J. *Measuring pupil achievement and aptitude*. (2nd ed.) New York: Harcourt Brace Jovanovich, 1975.

Lyman, H.B. *Test scores and what they mean*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

Mehrens, W.A., and Lehmann, I.J. *Measurement and evaluation in education and psychology*. New York: Holt, Rinehart and Winston, 1973.

Payne, D.A. *The assessment of learning: Cognitive and affective*. Lexington, MA: D.C. Heath, 1974.

Sax, G. *Principles of educational measurement and evaluation*. Belmont, CA: Wadsworth, 1974.

Stanley, J.C., and Hopkins, K.D. *Educational and psychological measurement and evaluation*. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Thorndike, R. L., & Hagen, E. P. *Measurement and evaluation in psychology and education*. (4th ed.) New York: Wiley, 1977.

Additional Measurement and Evaluation Textbooks

Blood, D. F., & Budd, W. C. *Educational measurement and evaluation*. New York: Harper and Row, 1972.

Dick, W., & Hagerty, N. *Topics in measurement: Reliability and validity*. New York: McGraw-Hill, 1971.

Gronlund, N. E. *Constructing achievement tests*. (2nd ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Hannah, L. S., & Michaelis, J. U. *A comprehensive framework for instructional objectives: A guide to systematic planning and evaluation*. Reading, MA: Addison-Wesley, 1977.

Kibler, R. J., Barker, L. L., & Miles, D. T. *Behavioral objectives*. Boston: Allyn & Bacon, 1970.

Mager, R. F., & Pipe, P. *Analyzing performance problems*. Belmont, CA: Fearon Publishers, 1970.

Nelson, C. H. *Measurement and evaluation in the classroom*. Toronto: The Macmillan Company, 1970.

Townsend, E. A., & Burke, P. J. *Using statistics in classroom instruction*. New York: Macmillan, 1975.

Martuza, V. P. *Applying norm-referenced and criterion-referenced measurement in education*. Boston: Allyn & Bacon, 1977.

Popham, W. J. *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Unit 3

Assessment of Content Validity

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

Table of Contents

Page

3.0 Overview of the Unit. 1

3.1 Introduction. 2

3.2 Judgments of Content Specialists. 6

3.2.1 Item-Objective Match. 6
A. An Index of Item Homogeneity 6
B. Semantic Differential Technique. 9
C. The Matching Procedure 12
D. Field Test of the Three Procedures 16
E. Item Review Form 17
F. Summary. 18
3.2.2 Representativeness of the Test Items. 18

3.3 Collection and Analysis of Student Response Data. 25

3.3.1 Standard Item Indices 28
3.3.2 Item Change Statistic 31
3.3.3 Items as Measures of Single Objectives. 34

3.4 Additional Editing of the Test Items. 35

3.5 References. 36

3.5.1 References Cited. 36
3.5.2 Additional References 37

3.0 Overview of the Unit

This unit was prepared to introduce practitioners to methods for determining the content validity of a set of test items. Principally there are two methods: Involvement of content specialists and the collection and analysis of student response data. In a final section, the matter of item revisions based on available data is considered.

3.1 Introduction

Generally speaking, the quality of criterion-referenced test items can be determined by the extent to which they reflect, in terms of their content, the domainsfrom which they were derived. The problem here is one of item validation; unless one can say with a high degree of confidence that the items in a criterion-referenced test measure the intended instructional objectives, any use of the test score information will be questionable. Thus far, the possible use of item generation forms, amplified objectives, and domain specifications have been considered. When item generation rules are used, a high degree of confidence in terms of items measuring intended objectives is derived through the direct relationship set up between items and the domain. This might be called an a priori approach to item validity; the approach itself assures that the items are valid, or representative of the domains. When amplified objectives or domain specifications are utilized, the domain definitions are never really precise enough to assume, a priori, that the items are valid. Thus, the quality of the items, in a context independent from the process by which the items were generated, must be determined. This is an a posteriori approach to item validation, and the procedures to be discussed are designed to assess whether or not a direct relationship between an item and a domain or objective exists through analysis of data collected after items are written.

There are two general approaches that may be used to establish the (content) validity of criterion-referenced test items. The first approach, and the approach we feel holds the most merit, involves the judgments of test items by content specialists. The judgments that are made concern the extent of "match" between the test items and the domains they are designed to measure. Questions asked of content specialists about content validity of test items can be reduced to two important ones (Hambleton, 1978):

1. Is the format and content of an item appropriate to measure some part of the domain specification?
2. Do the items adequately sample a particular domain?

The second approach is to apply empirical techniques, in much the same way as empirical techniques are applied in norm-referenced test development. In fact, along with some recently developed empirical procedures for criterion-referenced tests, several norm-referenced test item statistics can (and should) be used. The problem is to ensure that these statistics are used and interpreted correctly in the context of criterion-referenced test development. There are at least four problems involved with the use of empirical procedures. These problems are:

1. Most (if not all) of the procedures are dependent upon the characteristics of the group of examinees and the effects of instruction.
2. They often require sophisticated techniques and/or computer programs which are not available to practitioners.
3. When item statistics derived from empirical analyses of test data are used to "select" the items for a criterion-referenced test, the test developer runs the risk of obtaining a non-representative set of items from the domains measuring the objectives included in the test.
4. Empirical methods in many instances require pre-test and post-test data on the same items. Pretest data are rarely collected nor can they be. One reason is that there is a reluctance to administer tests to examinees where there is little change of moderate or high levels of performance.

Several criterion-referenced test theorists do espouse the use of empirical procedures for validating test items. However, one point to be made in this unit is that empirical procedures are less useful than the ratings of content specialists in the item validation process. It is often argued that many item statistics will be low (item discrimination indices, for example) because test score variance will be low, and therefore they will be of limited usefulness. On the contrary, authors such as Haladyna (1974) have found that there is usually sufficient test score variance. Also, test score variance can be assured by the selection of a proper item pilot study sample. Both "masters" and "non-masters" of the content under study should be located and included in a pilot sample. The fact is that empirical data is not very useful for answering the two content validity questions introduced earlier; and therefore empirical methods have limited usefulness. On the other hand, when construct validity evidence is being sought, examinee response data is exactly what is needed. However, empirical methods do have one important use in the content validation process. According to Rovinelli and Hambleton (1977):

> In situations where a large sample of examinees is available and where the test constructor is interested in identifying aberrant items, not for elimination from the item pool but for correction, the use of an empirical approach to item validation should provide important information with regard to the assessment of item validity.

In sum, the use of content specialists' ratings is the method to use for content validating test items; empirical procedures should be used only for the detection of aberrant items in need of correction. Unlike empirical procedures, the use of content specialists' ratings is not dependent on examinee group composition or instructional effects, may not require sophisticated statistical techniques, is not restricted to highly structured content domains, and finally, can be implemented easily in practical settings.

3.2 Judgments of Content Specialists

Content specialists must address two questions in assessing content validity:

1. Is the format and content of an item appropriate to measure some part of the domain specification?
2. Do the items adequately sample a particular domain?

Methods are offered in Sections 3.2.1 and 3.2.2 for addressing each question.

3.2.1 Item-Objective Match

A. An Index of Item Homogeneity

This technique is based upon the original work of Hemphill and Westie (1950) in constructing personality tests. The mechanism for collecting data consists of having content specialists rate each item on each of a set of objectives by assigning a value of +1, 0, or -1. These three possible ratings have the following meanings:

+1 = defining feeling that an item is a measure of the objective

0 = undecided about whether the item is a measure of the objective

-1 = definite feeling that an item is not a measure of the objective.

Basically, the content specialist's task is to make a judgment about whether or not an item is reflective of the content defined by a domain specification. If, for example, there are 10 objectives and 30 test items, each content specialist is required to make 300 judgments.

Rovinelli and Hambleton (1977), extended the work of Hemphill and Westie by developing a new statistic for providing a numerical representation of the data. They called this new statistic, the

Index of Item-Objective Congruence. The assumptions under which this index was developed are:

1. That perfect item objective congruence should be represented by a value of +1 and will occur when all the specialists assign a +1 to the item for the appropriate objective and a -1 to the item for all the other objectives.
2. That the worst value of the index an item can receive should be represented by a value of -1 and will occur when all the specialists assign a -1 to the item for the appropriate objective and a +1 to the item for all the other objectives.
3. That the value of the index should not depend on the number of content specialists or the number of objectives.

The index of item-objective congruence is given by

$$I_{ik} = \frac{(N-1) \sum_{j=1}^{n} X_{ijk} - \sum_{m=1}^{N} \sum_{j=1}^{n} X_{mjk} + \sum_{j=1}^{n} X_{ijk}}{2\,(N-1)n}$$

where

I_{ik} is the index of item-objective congruence for item k on objective i,

N is the number of objectives ($i=1, 2, \ldots, N$),

n is the number of content specialists ($j=1, 2, \ldots, n$),

X_{ijk} is the rating (-1, 0, +1) of item k as a measure of objective i by content specialist j.

The choice of a cut-off score to separate "valid" from "non-valid" items with the index should be based on experience with content specialists' ratings and with the index itself. In our work, when we feel it desirable to set a cutting score, we create the poorest set of content specialists' ratings that we would be willing to accept as evidence for the content validity of a test item. The value of the index for this set of minimally acceptable ratings serves as the cutting score for judging the item-objective match for each of the test items. For example, suppose that we have 20 content specialists and 10 objectives. We might desire that at least 15 of the content specialists match the item to the intended objective and that they indicate that the item is not a measure of the other nine objectives. In this case:

$$I_{ik} = \frac{9(15) - [(-9)(15) + (+1)(15)] + (15)}{2(9)(20)}$$

$$= \frac{135 - [-135 + 15] + 15}{360}$$

$$= \frac{135 + 120 + 15}{360}$$

$$= \frac{270}{360}$$

$$= .75$$

Note that for this example, N = 10, n = 20. The middle term in the numerator indicates how the judges (we want at least 15 of them to match the item to the intended objective and indicate lack of match to the other nine objectives) scored the item on all ten objectives. That is, the 15 gave a score of -1 on nine objectives and a score of +1 on the other, the intended objective. The final term corrects the bias built into the middle term by adding back into the numerator the scores subtracted out on the middle term for the intended objective. The value, .75, serves as the criterion against which item validities from the content specialists' ratings are judged.

The one major drawback of the approach is that it is very time consuming. Even if content specialists are assigned only a portion of the domain specifications and test items to review, the time required to rate the quality of each of a set of test items against all other domain specifications presented in a set can be substantial. Still, the approach is especially useful if there is reason to believe that test items may be measuring several objectives simultaneously.

B. Semantic Differential Technique

This technique employs the use of the semantic differential procedure (Osgood, Suci, and Tannebaum, 1957). Content specialists are presented with an objective and all the items on which ratings are desired. They are asked to make a judgment which consists of deciding whether the item-objective relationship is best described by the adjective toward the left-end or the right-end of the scale.

The following is an example consisting of one objective, one item, and two adjective scales, along with a set of typical directions:

Objective: Given the chemical formula for a molecule, determine the number of atoms in a molecule.

Item 1: How many atoms are there in a molecule of sulfuric acid H_2SO_4?

Directions

Given the objective and item above, your task is to make judgments on the relationship between the objective and the item on the adjective scales indicated below.

Scale 1:	very relevant	relevant	no feeling	irrelevant	very irrelevant
Scale 2:	very suitable	suitable	no feeling	unsuitable	very unsuitable

The data obtained from the use of this technique (more adjective scales would be desirable) can be analyzed without employing elaborate statistical techniques. Therefore, it can easily be used in practical settings. The information which is needed is the average scale score for each item on each objective rated by the content specialists. However, the data also lends itself to more elaborate statistical analysis. An examination of the standard deviations of the ratings given each item on each of the scales will provide an indication of the extent of agreement among the content specialists. An average across items and scales will give a general indication of the extent of the specialists' agreement. For instance, in a study done by Rovinelli and Hambleton where there were 48 items and a 5 point rating scale, the average standard deviation was .46. On the level of the item, with the exception of a few items, the standard deviations were quite small, thus indicating substantial agreement amont the content specialists' ratings.

A second procedure for assessing item-objective match involving the use of a rating scale was offered by Hambleton (1978). In this procedure, content specialists are given objectives (or domain specifications) and a set of test items. Their task is to rate the quality of test items as measures of the intended objectives (or domain specifications). A copy of a judge's rating form is presented in Figure 3.2.1.

Again, the rating scale data may be analyzed without employing any elaborate statistical procedures. It can easily be used in practical settings such as in the classroom by teachers. The information needed is the mean and median rating assigned by a group of content specialists to the items. An examination of the range of the ratings

Item Rating Form

Reviewer:_______________________ Date:_____________ Content Area:________

First, read carefully through the lists of domain specifications and test items. Next, please indicate how well you feel each item reflects the domain specification it was written to measure. Please use the five-point rating scale shown below:

Poor	Fair	Good	Very Good	Excellent
1	2	3	4	5

Circle the number corresponding to your rating beside the test item number.

Objective	Test Item	Item Rating					Comments
1	2	1	2	3	4	5	
	7	1	2	3	4	5	
	14	1	2	3	4	5	
2	1	1	2	3	4	5	
	3	1	2	3	4	5	
	8	1	2	3	4	5	
	13	1	2	3	4	5	
3	4	1	2	3	4	5	
	6	1	2	3	4	5	
	12	1	2	3	4	5	
4	5	1	2	3	4	5	
	9	1	2	3	4	5	
	10	1	2	3	4	5	
	11	1	2	3	4	5	

Figure 3.2.1 An example of a judge's item rating form.

given each item provides an indication of the extent of agreement among the content specialists.

It is also possible to determine the "closeness" of each judge's ratings to the median responses of the group. In some cases, when one or more of the judges are "far out-of-line" it may be best to eliminate their responses and recalculate the statistics. A summary and analysis of the hypothetical ratings of nine judges to 14 test items measuring four objectives is shown in Table 3.2.1.

C. The Matching Procedure

A third procedure which can be used to obtain the judgments of content specialists involves the use of a matching task. Content specialists are presented with two lists: One with test items and another with objectives (or domain specifications). The specialist's task is to indicate which objective he/she thinks each test item measures, if any. A contingency table is then constructed by calculating the numbers of content specialists matching each item to each objective in the sets of items and objectives being studied. The chi-square test for independence can then be used to analyze the data which is presented in the contingency table. Also, a simple visual analysis of the contingency table will reveal the amount of agreement among the specialists, and the types and location of disagreements. An example of a judge's set of directions for matching test items and objectives is presented in Figure 3.2.2. Some hypothetical results are reported in Table 3.2.2.

Table 3.2.1

Summary and Analysis of Judges' Ratings of 14 Test Items

		Judges' Ratings									Summary Statistics		
Objective	Test Item	1	2	3	4	5	6	7	8	9	Mean	Median	Rang
1	2	4	3	5	5	4	5	5	5	4	4.4	5	3
	7	4	2	5	5	5	5	5	4	5	4.4	5	4
	14	4	5	5	5	4	5	5	5	5	4.8	5	2
2	1	3	5	3	2	1	4	5	2	4	3.2	3	5
	3	3	1	4	4	3	4	4	3	3	3.2	3	4
	8	1	3	1	2	1	1	1	1	1	1.3	1	3
	13	1	3	2	1	1	2	1	2	3	1.8	2	3
3	4	4	5	5	4	5	5	5	5	5	4.8	5	2
	6	4	2	4	4	4	4	4	4	4	3.8	4	3
	12	5	3	5	5	5	5	5	5	5	4.8	5	3
4	5	4	3	5	5	4	5	5	4	5	4.4	5	3
	9	2	2	4	1	4	2	4	4	4	3.0	4	4
	10	1	3	1	2	1	1	1	1	1	1.3	1	3
	11	4	3	4	4	5	5	5	5	5	4.6	5	3
Judges' Discrepancies From Median Responses		9	24	1	10	6	4	4	3	3			

BEST COPY AVAILABLE

Items/Objectives Matching Task

Reviewer:________________________ Date:_______________ Content Area:_________

First, read carefully through the lists of domain specifications and test items. Your task is to indicate whether or not you feel each test item is a measure of one of the domain specifications. It is, if you feel examinee performance on the test item would provide an indication of an examinee's level of performance in a pool of test items measuring the domain specification. Beside each objective, write in the test item numbers corresponding the test items which you feel measure the objective. In some instances, you may feel items do not measure any of the available domain specifications. Write these test item numbers in the space provided at the bottom of the rating form.

Objective	Matching Test Items
1	
2	
3	
4	
No Matches	

Figure 3.2.2 An example of a judge's summary sheet for the items/objectives matching task.

Table 3.2.2

Summary and Analysis of the Judges' Item/Objective Matching Task

Objective	Test Item	Judges' Matches 1	2	3	4	5	6	7	8	9	Percent of Matches
1	2	1	0	1	0	1	1	1	1	1	78
	7	1	1	1	0	1	1	1	1	1	89
	14	0	1	1	1	1	1	1	1	1	89
2	1	0	0	1	1	1	1	1	1	1	78
	3	1	1	1	1	1	1	1	1	0	89
	8	1	0	1	0	0	0	0	1	0	33
	13	0	0	1	0	0	0	0	0	0	11
3	4	1	0	1	1	1	1	1	1	1	89
	6	0	0	1	0	1	0	1	1	0	44
	12	1	1	1	1	1	1	1	1	1	100
4	5	1	0	1	1	1	1	1	1	1	89
	9	1	1	1	1	0	0	1	1	1	78
	10	0	0	1	0	0	0	0	0	0	11
	11	1	1	1	1	1	1	1	1	0	89
Percentage of Matches for Each Judge		64	43	100	64	71	64	79	86	57	
"Lemons"	1	0	1	0	0	1	0	0	0	0	
	2	1	1	0	1	0	0	0	0	0	
	3	1	1	0	0	0	0	0	0	0	
Number of "Lemons" Misidentified		2	3	0	1	1	0	0	0	0	

Further, the "accuracy" of each content specialist can be checked if a specified number of "lemon" items (items not measuring any of the objectives) are introduced into the matching task. A content specialist's effectiveness can be measured by the number of "lemon" items detected. (Hambleton [1978] noted however that such a method of evaluation would not detect a "poor" judge if he/she was very critical of many of the test items.) Content specialists who fall short of some standard of performance can have their ratings removed from the statistical analysis of item ratings. One example of standard might be: A content specialist must identify correctly at least 75% of the "lemon" items.

D. Field Test of the Three Procedures

Rovinelli and Hambleton (1977) conducted an empirical study of the use of three procedures (test items matched to objectives using a three-point rating scale, semantic differential scale, and a matching task) using 48 items and 12 objectives from a ninth grade science curriculum. The reader is asked to refer to the article for details; we will summarize their findings here. They found that all three methods did provide useful information for ascertaining if an item is a measure of an objective. They also found some differences in the sorts of data collected through the use of the three techniques. For instance, the data appeared to show that the content specialists, when using the rating procedure (semantic differential), judged the items to be relevant measures of objectives other than the intended ones more often than when using the matching procedure. They recommend:

> Given the task of judging which items are measures of the intended objectives, the Hemphill-Westie procedure is recommended over the other two techniques. Two statements are offered in support of this recommendation. One, the numeric representation of the data, the index of item-objective congruence, provides a meaningful interpretation of the extent to which an item is judged to be a valid measure of the intended objective. Two, there are methods for determining the reliability and validity of the data collected. Further, these methods can be tested for statistical significance.

Rovinelli and Hambleton offer three cautionary notes about the use of the Hemphill-Westie procedure. One, the procedure does not give information on the quality of the items or the suitability of the distractors. Two, the dimensionality of the data must be known in advance. Three, the procedure is very time consuming for a large number of items and objectives. In sum, when deciding upon which of the procedures to use with content specialists' ratings, Rovinelli and Hambleton suggest:

> . . .before selecting the type of judgmental procedure to use, the test constructor should take into consideration the information desired and the resources available, and then choose the most appropriate procedure.

E. Item Review Form

An item review form for multiple-choice test items in the Reading/Language Arts area was introduced in Unit 2. The form appears again in Figure 3.2.3. A summary of the item reviews of several content specialists is very useful in the content validation process.

Recently, another item review form and instructions for its use were developed and piloted in a multiple-choice item writing workshop conducted in Baton Rouge, Louisiana. The material is presented in Figure 3.2.4. The special feature of this new form is that the item ratings for up to ten test items measuring an objective can be reported on a single page.

Forms in Figures 3.2.3 and 3.2.4 can easily be prepared for other item formats using the item writing principles offered in Unit 2.

F. Summary

Data from all of the procedures sketched out in Section 3.1.1 are useful for determining the content validity of a set of test items. The data derived from any one of the procedures can be used to answer the following important questions:

1. Which items failed to "match" the domain specifications they were prepared to measure?
2. How successful were the test item writers?
3. How can the content validity data on the test items be used to rewrite domain specifications?
4. Who were the "best" content specialists in the rating process?

3.2.2 Representativeness of the Test Items

This step cannot be completed until the test items to be included in a test have been selected. It is usually desirable to have test items in a criterion-referenced test that are representative of the domain of items indicated in a domain specification, i.e., criterion-referenced tests

Figure 3.2.3 A sample multiple-choice item review form.

Item Review Form
(Multiple-Choice)

Objective Number: ______________ Test Item Number: ______________

Reviewer: ____

Date: _________________________

Objective:

Test Item:

Section I. Technical Quality

Place a "✓" under the column corresponding to your rating of the test item for the questions in this section and the next one.

	Yes	Questionable	No
1. Is the item stem clearly written for the intended group of examinees?	___	___	___
2. Is the item stem free of irrelevant material?	___	___	___
3. Is a problem clearly defined in the item stem?	___	___	___
4. Are the choices clearly written for the intended group of examinees?	___	___	___
5. Are the choices free of irrelevant material?	___	___	___

	Yes	Questionable	No
6. Is there a correct answer or a clearly best answer?	___	___	___
7. Have words like "always," "none," or "all" been removed?	___	___	___
8. Are likely examinee mistakes used to prepare incorrect answers?	___	___	___
9. Is "all of the above" avoided as a choice?	___	___	___
10. Are the choices arranged in a logical sequence (if one exists)?	___	___	___
11. Was the correct answer randomly positioned among the available choices?	___	___	___
12. Are all repetitious words or expressions removed from the choices and included in the item stem?	___	___	___
13. Are all of the choices of approximately the same length?	___	___	___
14. Do the item stem and choices follow standard rules of punctuation and grammar?	___	___	___
15. Are all negatives underlined?	___	___	___
16. Are grammatical cues between the item stem and the choices which might give the correct answer away removed?	___	___	___
17. Is the item format appropriate for measuring the intended objective?	___	___	___

Section II. Technical Quality Matters Specific to Reading/Language Arts Test Items

	Yes	Questionable	No
1. Can a correct answer be given without reading the passage?	___	___	___
2. Is the discourse appropriate for measuring the intended objective?	___	___	___

	Yes	Questionable	No
3. Does the discourse and test item provide a valid measure of the intended objective?	___	___	___
4. Do the following fall within the range for the number of words in each sentence of the			
(a) directions?	___	___	___
(b) discourse?	___	___	___
(c) item stem?	___	___	___
(d) item choices?	___	___	___
5. Do the following fall within the range for the number of sentences in the			
(a) directions?	___	___	___
(b) discourse?	___	___	___
(c) item stem?	___	___	___
(d) item choices?	___	___	___
6. Is there the desired number of words in the selection of discourse?	___	___	___
7. Does the test item contain the desired number of choices?	___	___	___
8. Is the ratio of common to uncommon words correct?	___	___	___

- -

Suggested Revisions:

- -

Final Rating (Check One):

☐ Accept

☐ Accept (with revisions-see above)

☐ Reject

Figure 3.2.4 Instructions for Using the Item Review Form

1. Obtain a copy of a domain specification and the test items written to measure it.

2. Place the domain specification number, your name, and today's date in the space provided at the top of the Item Review Form.

3. Place the numbers corresponding to the test items you will evaluate in the spaces provided near the top of the Item Review Form. The numbers should be in ascending order as you read from left to right. (This must be done if processing of your data along with data from many other reviewers is to be done quickly and with a minimum number of errors.)

4. Read the domain specification carefully.

5. Read the first test item carefully and answer the first 18 questions. Mark "✓" for "yes"; mark "X" for "No"; and mark "?" if you are "unsure."

   The last question requires you to provide an overall evaluation of the test item as an indicator of the domain specification it was written to measure.

   There are five possible ratings:

   5 - Excellent
   4 - Very Good
   3 - Good
   2 - Fair
   1 - Poor

6. Write any comments or suggested wording changes on or beside the test item.

7. Repeat the rating task for each of the available test items.

8. Staple your Item Review Form, domain specification, and copy of the test items together.

Item Review Form
Multiple Choice

Domain Specification No. ______ Reviewer ______________________________ Date ________________

Test Item Characteristics (Mark "√" for Yes, "X" for No, and "?" for Unsure)	Test Item Numbers									
1. Is the item stem clearly written for the intended group of students?										
2. Is the item stem free of irrelevant material?										
3. Is a single problem clearly defined in the item stem?										
4. Are the answer choices clearly written for the intended group of students?										
5. Are the answer choices free of irrelevant material?										
6. Is there a correct answer or a clearly best answer?										
7. Have words like "always," "none," or "all" been removed?										
8. Are likely student mistakes used to prepare incorrect answers?										
9. Is "all of the above" avoided as an answer choice?										
10. Are the answer choices arranged in a logical sequence (if one exists)?										
11. Was the correct answer randomly positioned among the available answer choices?										
12. Are all repetitious words or expressions removed from the answer choices and included in the item stem?										
13. Are all of the answer choices of approximately the same length?										
14. Do the item stem and answer choices follow standard rules of punctuation and grammar?										
15. Are all negatives underlined?										
16. Are grammatical cues between the item stem and the answer choices, which might give the correct answer away, removed?										
17. Are letters used in front of the possible answer choices to identify them?										
18. Have expressions like "which of the following is not" been avoided?										
19. Disregarding any technical flaws which may exist in the test item (addressed by the first 18 questions), how well do you think the content of the test item matches with some part of the content defined by the domain specification? (Remember the possible ratings: 1=poor, 2=fair, 3=good, 4=very good, 5=excellent)										

must be content valid. Only in some highly special cases has it been possible to completely specify a pool of relevant test items. For example, there have been some successes in the areas of mathematics and spelling. But these examples are far removed from the content worlds of interpretative poetry, creative writing, and finite projective geometry. What then is to be done? Should we "close up shop" and fade back into the murky interpretative world of norm-referenced testing?

For starters, test developers need to work hard to define and to develop their domain specifications. If content issues are clarified fully enough, content specialists can comment on the apparent representativeness of items included in a test. An even better procedure is Cronbach's duplication experiment. The experiment requires two teams of equally competent item writers and reviewers to work independently in developing a criterion-referenced test. Cronbach's (1971) directions are:

> They would be aided by the same definition of relevant content, sampling rules, instructions to reviewers, and specifications for tryout and interpretation of the data. . .

If the domain specification is clear, and if sampling is representative, the two tests should be equivalent. We could check this by administering both tests to the same group of examinees. One problem is that "a common blind spot is almost impossible to detect" (Cronbach, 1971).

3.3 Collection and Analysis of Student Response Data

While test score variability is not a factor in criterion-referenced test construction, neither is it a completely useless concept. Indeed, variability will be observed when a sample of examinees is heterogenous in terms of their domain scores. By establishing a priori the composition of an examinee sample, the resulting variability will provide additional, helpful information for assessing test items. Haladyna (1974) offered a procedure for circumventing the problem of lack of variance in criterion-referenced test scores, thus allowing the use of traditional (norm-referenced) item discrimination indices. If there are two samples of students, one sample instructed on the objectives comprising the test and another group uninstructed, (or groups of "masters" and "non-masters" after instruction), these samples can be combined, thus increasing the variability in the scores to the extent that a traditional point-biserial correlation coefficient (an index of item discrimination) can be utilized.

Item statistics, such as discrimination indices (Cox and Vargas, 1966; Crehan, 1974; Haladyna, 1974; Henrysson and Wedman, 1974), may provide useful information for detecting "bad" items. Indeed, Wedman (1973) gives a compelling argument for using item statistics. He argues that even carefully prepared domain specifications and precise item generation specifications never completely eliminate the subjective judgments that, to greater and lesser degrees, influence the test construction process. In order to guard against this subjective element, albeit small, domain specifications and item generating procedures should be complemented with empirical evidence on the items.

Essentially, empirical procedures involve the use of various item statistics that measure item difficulty and item discrimination. In most instances, for these statistics to be meaningful, it is necessary to have some item variability across examinees.

There has been discussion on the matter of item and test variance with criterion-referenced tests (Haladyna, 1974; Millman and Popham, 1974; Woodson, 1974a,b). Our own view, which is in agreement with Millman and Popham (1974), is that item and test variance are unnecessary with criterion-referenced tests. It is important that a criterion-referenced test have content validity and scores derived from the test must have construct validity (i.e., the test must measure what we say the test measures). Construct validity can be assessed in several ways (this point will be discussed more fully in Unit 5). For example, some variability of estimated domain scores could be expected across a pool of examinees consisting of "masters" and "non-masters" and to the extent that there was no (or limited) variability, the construct validity of the test scores (assuming content validity had been established) should be questioned. The test ought to reflect some variability of scores across "masters" and "non-masters" groups (perhaps post- and pre-instruction groups) although one would not select items to maximize the difference between scores in the two groups since that would make it difficult to obtain "valid" estimates of domain scores.

A point that must be stressed here is this: Item statistics derived from a field test should not serve as the sole criterion for refining an item pool or used to construct a criterion-referenced test. As Millman (1974) noted, "Item statistics can, however, be used to detect flawed items" (p. 339).

In discussing the various criterion-referenced test item statistics, a bit of background information may be helpful. When criterion-referenced testing received its "birth" in the late sixties and early seventies, the favored inclination was to try to pattern procedures for criterion-referenced tests after those of the already well-developed norm-referenced procedures. For norm-referenced tests, the item indices, item difficulty and item discrimination, were helpful in making decisions about test items. Naturally, the inclination of criterion-referenced theorists was to try to apply these indices, especially item discrimination indices (such as the point-biserial correlation coefficient) to criterion-referenced test items. The problem with such an approach is that these indices are built upon the concept of correlation, and correlation analysis is dependent upon a degree of variability in the data. This is not likely to be the case in criterion-referenced testing situations where most of the students should achieve mastery of the behavior in question, and thus the test scores will have little variability. Quite simply, the norm-referenced indices are consonant with the implicit use of norm-referenced tests, to facilitate comparisons of students, and not with the use of criterion-referenced tests, to indicate how much a student knows.

Various writers in the criterion-referenced testing field have developed indices for detecting aberrant test items and these indices don't suffer the problem of norm-referenced indices applied in CRT situations, but there is little agreement about which is optimal for a given situation. Because of this fact, and because of the following three points we offer, we feel that these indices should be used with a great deal of caution. In particular, we feel the indices should be used to detect aberrant items that need to be reworked, and not to make decisions about which items should and should not be on the test. The three points we make are:

1. The methods are based on the performance of a specific group of examinees, and thus this greatly limits the generalizability of the results.

2. It is difficult to determine the impact of instruction on these item statistics.

3. Many of the procedures require pre-test and post-test data on the same set of test items. This data is not likely to be collected by practitioners in classroom settings.

Next, several of the more promising item statistics will be considered. For an excellent review of these and other statistics, the interested reader is referred to a paper by Berk (1978).

3.3.1 Standard Item Indices

There are a number of standard statistical indices which appear to provide useful information for determining whether the items are adequate measures of the instructional objectives they were written to measure. When items in a domain are expected to be relatively homogeneous (this would be the case if the domain is defined narrowly), it has become a fairly common practice for the test developer to compare estimates of item difficulty parameters, or item discrimination parameters, or both. A typical item analysis printout for three items is shown in Figure 3.3.1. Since one would expect items measuring an objective equally well to have similar item parameters, estimates of the parameters are compared to detect items that deviate from the norm defined by the remaining items. Such "deviant" items are carefully scrutinized. In particular, content specialists' judgments of the "deviant" items are considered. If the items look acceptable, they are returned to the item domain. (Of

Figure 3.3.1 A computer print-out of a standard item analysis

```
ITEM      PER CENT CORRECT      PER CENT CORRECT      RBISERIAL WITH
NUMBER    FOR STUDENTS WHO      FOR ALL STUDENTS       TOTAL SCORE
            ATTEMPTED ITEM
 15           0.5758                0.5758               0.3999

                    NUMBER OF STUDENTS ANSWERING
                   EACH ALTERNATIVE BY QUARTER

QUARTER     NOT OMITTED      1      2      3      4      5      TOTAL

          REACHED
1            0       0       0      4     12      0      0        16
2            0       0       1      4     10      1      0        16
3            0       0       2      1     11      1      2        17
4            0       0       2      6      5      2      2        17
5            0       0       5     15     38      4      4        66

ITEM      PER CENT CORRECT      PER CENT CORRECT      RBISERIAL WITH
NUMBER    FOR STUDENTS WHO      FOR ALL STUDENTS       TOTAL SCORE
            ATTEMPTED ITEM
 20           0.6364                0.6364               0.3954

                    NUMBER OF STUDENTS ANSWERING
                   EACH ALTERNATIVE BY QUARTER

QUARTER     NOT OMITTED      1      2      3      4      5      TOTAL
          REACHED
1            0       0       0      0      0      0     16        16
2            0       0       7      0      0      0      9        16
3            0       0       4      0      3      0     10        17
4            0       1       4      0      5      0      7        17
5            0       1      15      0      8      0     42        66

ITEM      PER CENT CORRECT      PER CENT CORRECT      RBISERIAL WITH
NUMBER    FOR STUDENTS WHO      FOR ALL STUDENTS       TOTAL SCORE
            ATTEMPTED ITEM
 39           0.3788                0.3788               0.4254

                    NUMBER OF STUDENTS ANSWERING
                   EACH ALTERNATIVE BY QUARTER

QUARTER     NOT OMITTED      1      2      3      4      5      TOTAL
          REACHED
1            0       0       2      3      0     11      0        16
2            0       0       3      4      1      7      1        16
3            0       0       3      6      1      4      3        17
4            0       0       8      3      1      3      2        17
5            0       0      16     16      3     25      6        66
```

course, it may also be the case that items sharing similar statistical properties still do not measure the intended objectives and the so called "deviant" items do! Hence, there is a need to check the empirical results with the content specialists' ratings.) A more formal method of comparing item difficulty parameters is considered next.

Brennan and Stolurow (1971) present a set of rules for identifying criterion-referenced test items which are in need of revision. The decision process which they established for deciding which items to revise can be used to help assess item validity. However, our particular interest is with their procedure for comparing difficulty levels of items intended to measure the same objective. Brennan and Stolurow (1971) state that the item scores from criterion-referenced tests will most likely not be normally distributed. Therefore, in order to determine if the item difficulties are equal, they propose the use of Cochran's Q test. This statistic can be used to determine whether two or more item difficulties differ significantly among themselves. Cochran's Q is a test of the hypothesis of equal correlated proportions. For a large enough sample of examinees, Q is approximately distributed as a χ^2 variable with K-1 degrees of freedom, where K is the number of test items. To reject the null hypothesis, however, provides no guidance as to which items are significantly different. If the null hypothesis is rejected, pair-wise comparisons need to be computed to locate deviant items.

Perhaps we should note here that since the major purpose of criterion-referenced tests is to provide information for describing individual levels of mastery, one should compare item difficulties of items intended to measure the same objectives and which have been administered to the same group of examinees either before or after instruction. While it would be possible to compare items administered to different groups receiving the same instruction, the assessment problem would become more complex. This complexity arises from the need to determine whether the group compositions were the same and whether the instruction was equally effective in each group. We note though that comparing item difficulties only makes sense as part of an item validation process when the domain of items spanning an objective is considered to be homogeneous. There are many times when this assumption will be untenable (Millman, 1974).

One would also expect the intercorrelations of items intended to measure the same objective to be equal for a group of items homogeneous with respect to that objective (Brennan and Stolurow, 1971). If the test developer is willing to assume that the departure from normality for scores on the items is not a crucial problem, then there is a technique available to test for the equality of pairs of product moment correlation coefficients. When this assumption is not tenable, test developers will have to make subjective judgments as to the equality of these inter-item correlation coefficients.

3.3.2 Item Change Statistic

The difference between the difficulty level of an item before and after instruction describes another item statistic that seems to have some usefulness in the validation of criterion-referenced test items. However, an important point to note is that a large difference between the pretest and posttest item difficulty is not necessary since items may be valid indicators of the desired objectives but because of poor instruction, there may be very little change in difficulty level between the two test administrations. On the other hand, if instruction is effective, one would expect to see a substantial change in item difficulty _if_ the item is a measure of the intended objective. With several items intended to measure the same objective, one could also compare the item change indices for the purpose of detecting items that seem to be operating differently from the others.

Cox and Vargas (1966) were the first to suggest the use of an index linked to the instructional process. Their posttest-pretest _difficulty index_ is obtained by computing the percentage of examinees who pass an item on the posttest minus the percentage who pass the item on the pretest. Cox and Vargas ranked items on the basis of this index and correlated these rankings with those obtained through the use of a traditional norm-referenced test item index (the percentage of students in the highest 27% in total posttest scores who pass the item minus the percentage of the lowest 27% who pass the item). The correlation between the two item statistics were sufficiently low to allow Cox in a later paper (1970) to note:

> The pretest and posttest method of item analysis produced results sufficiently different from traditional methods to warrant its consideration in those cases where score variability is not the concern, such as in criterion-referenced measures.

Popham (1971) proposed a priori and a posteriori approaches for developing valid criterion-referenced test items. The a priori approach corresponds to the determination of validity by operationally obtaining items from an item generation rule. The a posteriori approach consists of empirically determining whether or not items are defective. In his discussion of the a posteriori approach, Popham presented a new means for empirically evaluating criterion-referenced test items. This procedure represents an extension of the item change statistic and consists of constructing the following four-fold table from the results of a pre-posttest administration of a set of items measuring an objective:

		Posttest	
		Incorrect	Correct
Pretest	Incorrect	A	B
	Correct	C	D

A, B, C, and D represent the number of examinees obtaining each of the four possible response patterns for an item. One then computes the median value across items measuring the same objective for each of the four cells. (The median value is not as likely to be affected by aberrant items, as would the mean.) These values are used as expected values and a chi-square statistic is computed (with three degrees of freedom) for each item.

An alternate way of looking at this procedure is to consider the median values for the four cells across the items measuring a particular objective as constituting a "prototypic" item. Then we can contrast the actual four-fold frequencies for each item to the frequencies in the cells for this prototypic item; large values of the "contrasting" statistic, the chi-square, would indicate an atypical item.

Three comments can be made concerning this index. One, the index is a measure of homogeneity. Popham states that this procedure was more accurate than visual scanning in locating the atypical items. While Popham (1971) describes other descriptive statistics, the chi-square analysis for detecting "bad" items seems to be the most promising one that he offered. Comments two and three are limitations that need to be mentioned. Popham (1971) indicated that there is no established critical value for chi-square, above which the items can be regarded as aberrant. Three, Wedman (1973) has pointed out that the method could lead to the elimination of items the test developer thinks are aberrant, but it does not take into consideration the direction of the abnormality. If the bulk of the items were, at best, mediocre in terms of representing the domain, a good item could be eliminated.

3.3.3 Items as Measures of Single Objectives

The concern here is with determining whether or not items are measuring one objective. Davis and Diamond (1974) have defined the importance of this condition:

> Unless all the items in a test measure exactly the same variable or variables for which true scores are highly correlated (say, .90 or greater), it is inappropriate to use the test for diagnostic purposes; that is, to determine an examinee's level of performance on a single "pure variable." This is because of the fact that two different examinees may obtain identical scores by marking correctly the same number of different items. . .
>
> The implication for the preparation of homogeneous items for a multi-item diagnostic test is that each item must measure only one "pure" variable plus error or the same weighted combination of two or more "pure" variables, plus error. In either of these cases, the item scores would be found to measure, at a pre-selected level of significance, the same dimension except for errors of measurement and for differences of origin and of units of measurement. . .

In reference to testing whether a criterion-referenced test item measures more than one objective, factor analysis offers great untapped potential.

3.4 Additional Editing of the Test Items

One problem for further research involves the development of a method for the integrative use of content specialists' ratings and empirical methods. While such an integration of approaches could be accomplished through logical analysis, perhaps a better way to proceed would be to actually employ the different techniques, in a variety of situations and through this practical experience, evolve a model for the combined use of content specialists' ratings and empirical methods. The work by Cronbach (1971) may help to provide a conceptual framework for this integration since his treatment of test validity is the most comprehensive to-date. Much work remains to be done in this area.

Suffice to say, the test developer's task is to use the available empirical data from content specialists and examinees to determine whether in his/her best judgment the available data supports the hypothesis that the items are "valid" measures of the intended objectives. When the data suggests otherwise, every effort should be made to revise aberrant items.

3.5 References

3.5.1 References Cited

Berk, R. A. A consumer's guide to criterion-referenced test item statistics. *Measurement in Education*, 1978, *9*, 1-12.

Brennan, R. L., & Stolurow, L. M. An empirical decision process for formative evaluation. *Research Memorandum No. 4*. Harvard CAI Laboratory, Cambridge, Mass., 1971.

Cox, R. C. Evaluative aspects of criterion-referenced measurement. Paper presented at the annual meeting of AERA, Minneapolis. 1970. (ERIC, Ed 038 679).

Cox, R. C., & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1966.

Davis, F. B., & Diamond, J. J. The preparation of criterion-referenced tests. *CSE monograph series in evaluation*. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. *Journal of Educational Measurement*, 1974, *11*, 93-99.

Hambleton, R. K. Validation of criterion-referenced test score interpretations and standard setting methods. A paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C., 1978.

Hemphill, J., & Westie, C. M. The measurement of group dimensions. *Journal of Psychology*, 1950, *29*, 325-342.

Hively, E., Maxwell, G., Rabehl, G., Sension, D., & Lundin, S. *Domain-referenced curriculum evaluation: A technical handbook and a case study from the Minnemast Project*. CSE monograph series in evaluation, No. 1. Los Angeles: Center for the Study of Evaluation, University of California, 1973.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, California: McCutchan Publishing Co., 1974.

Millman, J., & Popham, W. J. The issue of item and test variance for criterion-referenced tests: A clarification. *Journal of Educational Measurement*, 1974, *11*, 137-138.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning*. Urbana, IL: University of Illinois Press, 1957.

Popham, W. J. Indices of adequacy for criterion-referenced test items. In W. J. Popham (Ed.), *Criterion-referenced measurement: An introduction*. Englewood Cliffs, N.J.: Educational Technology Publications, 1971.

Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Dutch Journal for Educational Research, 1977, 2, 49-60.

Wedman, I. On the evaluation of criterion-referenced tests. Paper presented at the International Symposium on Educational Testing, the Hague, the Netherlands, 1973.

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64. (a)

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests: A reply. Journal of Educational Measurement, 1974, 11, 139-140. (b)

3.5.2 Additional References

Berk, R. A. Criterion-referenced test item analysis and validation. A paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, D.C. 1978.

Crehan, K. D. Item analysis for teacher-made mastery tests. Journal of Educational Measurement, 1974, 11, 255-262.

Henrysson, S., & Wedman, I. Some problems in construction and evaluation of criterion-referenced tests. Scandinavian Journal of Educational Research, 1974, 18, 1-12.

Herbig, M. Item analysis by use in pre-tests and post-tests: A comparison of different coefficients. Programmed Learning and Educational Technology, 1976, 13, 49-54.

Unit 4

Test Assembly and Administration

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

Table of Contents

Page

4.0 Overview of the Unit. 1

4.1 Introduction. 2

4.2 Determination of Test Length. 3

- 4.2.1 Introduction. 3
- 4.2.2 The Basic Situation 4
- 4.2.3 Millman's Use of the Binomial Model 5
- 4.2.4 Novick and Lewis' Bayesian Approach: Introduction . . 10
  - A. Examples of the Effects of Different Priors. . . . 13
  - B. Losses Associated with Incorrect Decisions 17
  - C. Test Length Specifications 18
  - D. Suggestions. 25
- 4.2.5 Fhaner's and Wilcox's Use of Indifference Zones . . . 25
  - A. Summary of the Procedure 30
  - B. Comments 30
- 4.2.6 Eignor-Hambleton Approach for Determining Test Length 32
  - A. Introduction 32
  - B. Research Design. 35
  - C. Results and Discussion 45
  - D. Suggestions for Further Research and Development . 55
- 4.2.7 A Method of Selecting a Procedure for Determining Test Length. 58

4.3 Test Item Selection 59

- 4.3.1 Post Item Selection Checklist 63

4.4 Preparation of Directions 66

4.5 Layout and Test Booklet Preparation 69

4.6 Preparation of Scoring Keys 71

4.7 Preparation of Answer Sheets. 72

4.8 Test Administration 73

4.9 References Cited. 76

4.0 Overview of the Unit

This unit covers steps 7 and 9 of the Criterion-Referenced Test Development and Validation Model presented in Unit 1. These steps are:

7. Test Assembly
    - a. Determination of Test Length
    - b. Test Item Selection
    - c. Preparation of Directions
    - d. Layout and Test Booklet Preparation
    - e. Preparation of Scoring Keys
    - f. Preparation of Answer Sheets
9. Test Administration

Four procedures are offered in Section 4.2 for determining test length. The remainder of the material in the unit (covering steps 7b, . . ., 7f, and 9) is straightforward. Our discussion of these steps for criterion-referenced test development is very similar to the discussion one would find of these steps for preparing norm-referenced tests.

Note: It is likely that some of the material in Section 4.2 will be more meaningful if Units 5 and 6 are studied first.

4.1 Introduction

In Unit 4, we will discuss research and procedures directed toward the assembly and administration of a criterion-referenced test. In many of the sections we will offer checklists that can aid in the process. The material presented here will vary greatly in difficulty, and in length of presentation. For instance, a great amount can be presented about how to go about determining the number of test items per objective, while little can be said about the preparation of test directions. In the sections that duplicate established principles for norm-referenced tests, we have presented a synthesis of the research pertaining to the section, and have directed readers to the source or sources from which it came.

4.2 Determination of Test Length

4.2.1 Introduction

The length of a criterion-referenced test (or more importantly, the number of test items measuring each objective in a test) is directly related to the usefulness of the criterion-referenced test scores obtained from the test. Short tests, typically, produce imprecise domain score estimates, and lead to mastery decisions which prove to be inconsistent across parallel-form administrations (or retest administrations). Therefore, criterion-referenced test scores obtained from short tests have limited value. When estimation of domain scores is of concern, the relationships among domain scores, errors of measurement, and test length, as summarized in the item-sampling model, are well known (Lord and Novick, 1968) and provide a basis for determining test length.

When using criterion-referenced tests to assign examinees to mastery states, the problem of determining test length can be related to the number of classification errors one is willing to tolerate. One way to assure low probabilities of misclassification is to make the test very long. However, this is not usually feasible. Currently, there exist at least two ways to reduce classification errors without lengthening a test. One involves utilizing Bayesian estimation procedures incorporating prior and collateral information. The second involves the implementation of an adaptive testing scheme especially designed for hierarchically-structured objectives (see Hambleton & Eignor, 1978; Spineti & Hambleton, 1977).

The material and procedures to follow can be separated roughly into four sections. One section involves the work of Millman (1973), utilizing the binomial model. The second section involves the work of Novick and Lewis (1974), using Bayesian estimation procedures. The third section involves the specification of an "indifference zone." The work of Fhaner

(1974) and Wilcox (1976) will be considered here. The final section includes the work of Eignor and Hambleton (1979) relating test length and cut-off scores to several reliability and validity indices.

4.2.2 The Basic Situation Revisited

Regardless of which solution one adopts to the test length problem, the basic situation remains the same. It is as follows: There is a domain or population of test items (it may be real or have to be hypothesized). These items deal with a particular objective and are of varying unspecified difficulty. We want to pass the student on the objective if he/she can answer a given percentage of items in the domain. This actual percentage of items the student could pass on the whole domain or population of items can be called his/her domain score. Practical constraints, such as time, force the test practitioner to have to estimate this domain score by taking a random or stratified random sample of items from the domain and testing on those items. Now, because the test score is based on a sample from the domain, it is not likely to coincide with the domain score. There will be error, and from the point of view adopted for criterion-referenced tests, we view the error as error in the decision process. That is, the extent that an individual's test score is discrepant from his/her domain score can be viewed as a problem involving the probability of classifying that individual improperly, i.e., as a false positive (a non-master who is assessed as a master on the test) or a false negative (a master assessed as a non-master on the test). Logic dictates that the longer the test is, the less the chance of making classification errors. Practicality dictates against having long tests, due to time problems, construction problems, etc. Thus, the concern becomes one of determining what minimal test length is sufficient in terms of the problem of classification errors.

4.2.3 Millman's Use of the Binomial Model

Millman (1973) considered the error properties of mastery classification decisions made by comparing a domain score estimate to an advancement score. By introducing the binomial test model, it is simple to determine the probability of misclassification, conditional upon an examinee's domain score, an advancement score, a cut-off score, and the number of items in the test. (An advancement score is distinguished from a cut-off score in Millman's work in the following way: The advancement score is the minimum number of items that an examinee must answer correctly to be assigned to a mastery state. The cut-off score is the point on the domain score scale used to separate examinees into true mastery and true non-mastery states.) By varying test length and the advancement score, an investigator can determine the test length and advancement score that produces a desired probability of misclassification for a given domain score.

By making the following assumptions, Millman was able to obtain a solution to the test length problem:

1. The test is a random sample of dichotomously scored (0-1) items from the domain,
2. The likelihood of correct response is a fixed quantity across all test items for an individual,
3. Responses to questions on the test are independent, and
4. Errors fit the binomial test model.

No assumptions involving item content or difficulty are necessary, nor are any group based indices used. Millman (1973) compared the situation to the usual urn example used for explaining the binomial distribution:

> Rather the items which an examinee can pass and
> those the individual fails are analogous to two
> colors of balls in an urn. Continuing the analogy,

> the test length question, How many balls must be sampled (items administered) from the urn so that the percent of all balls of a given color (test items in the domain answered correctly) can be estimated accurately? The errors associated with other examinees are of no concern.

Table 4.2.1 can be used to obtain the probability that a student with a particular domain score will be incorrectly advanced or retained by the procedures. (It is assumed that some meaningful method has been utilized to arrive at the cut-off score.) The following comments can be made concerning the use of Table 4.2.1:

1. To the left of the dotted line indicating the cut-off score is the expected percent of students who will be advanced incorrectly. Likewise, to the right of the dotted line is the expected percent of students who will be incorrectly retained. Their domain scores are greater than the criterion level. In other words, to the left of the line are the false-positive error rates and to the right are the false-negative error rates.

2. A larger proportion of the students whose domain scores are close to, or at the cut-off score, will be incorrectly classified than those at a greater distance from the cut-off score. Sometimes this proportion is greater than half. For instance, for a cut-off score = .75 on a test with 8 items that has an advancement score of 6, a student whose domain score is 70 will be incorrectly advanced (passed) 55% of the time.

3. Millman looks at the probability that a student will attain a particular test score, given his/her domain score. However, an examinee's domain score is an unknown. It is, of course, the purpose of testing in the first place! On the other hand, it is usually not too difficult, in most situations, to make an educated guess.

Example 1:

For a cut-off score of .80, suppose a practitioner is willing to accept a 25% misclassification error for those students whose domain scores are 70% and 90%. How large should the random sample of items be,

Table 4.2.1 Percent of Students Expected to be Incorrectly Advanced or Retained

Cut-off Score = .70

Advancement Score	No. of Test Items	Student's Domain Score* 50	55	60	65	70	75	80	85	90	95
6	7	6	10	16	23	67	55	42	28	15	4
6	8	15	22	32	43	45	32	20	11	4	1
7	9	9	15	23	34	54	40	26	14	5	1
7	10	17	27	38	51	35	22	12	5	1	-
8	11	11	19	30	43	43	29	16	7	2	-
9	12	7	13	23	35	51	35	20	9	3	-
10	13	5	9	17	28	58	42	25	12	3	-
11	14	3	6	12	22	64	48	30	15	4	-
12	15	2	4	9	17	70	54	35	18	6	-

Cut-off Score = .75

Advancement Score	No. of Test Items	Student's Domain Score* 50	55	60	65	70	75	80	85	90	95
6	8	15	22	32	43	55	32	20	11	4	1
7	9	9	15	23	34	46	40	26	14	5	1
8	10	6	10	17	26	38	47	32	18	7	1
9	11	3	7	12	20	31	55	38	22	9	2
9	12	7	13	23	35	49	35	20	9	3	-
16	20	1	2	5	12	24	58	37	17	4	-
17	21	-	1	4	9	20	63	41	20	5	-
18	22	-	1	3	7	17	68	46	23	6	-

Table 1 (continued)

Cut-off Score = .80

Advancement Score	No. of Test Items	Student's Domain Score* 50	55	60	65	70	75	80	85	90	95
6	7	6	10	16	23	33	45	42	28	15	4
7	8	4	7	11	17	26	37	50	34	19	6
8	9	2	4	7	12	20	30	56	40	23	7
8	10	6	10	17	26	38	53	32	18	7	1
9	11	3	7	12	20	31	46	38	22	9	2
10	12	2	4	8	15	25	39	44	26	11	2
11	13	1	3	6	11	20	33	50	31	13	2
12	15	2	4	9	17	30	46	35	18	6	-
17	20	-	1	2	4	11	23	59	35	13	2
19	22	-	-	1	3	7	16	67	42	17	2

Cut-off Score = .85

Advancement Score	No. of Test Items	Student's Domain Score* 50	55	60	65	70	75	80	85	90	95
7	8	4	7	11	17	26	37	50	34	19	6
8	9	2	4	7	12	20	30	44	40	23	7
9	10	1	2	5	9	15	24	38	46	26	9
10	11	1	1	3	6	11	20	32	51	30	10
11	12	-	1	2	4	9	16	28	56	34	12
17	19	-	-	1	2	5	11	24	56	29	7
19	21	-	-	-	1	3	8	18	63	35	8

*A domain score is the proportion of items a student would be able to answer correctly if he/she were given the entire pool of items measuring an objective.

(Reproduced from Novick and Lewis, 1974, with permission from the authors. Decimal points have been omitted.)

and what should the advancement score be?

Answer: From Table 1, it can be seen that for a test of 8 items with a passing (advancement) score of 7, 26% of the students at the 70% level and 19% at the 90% level will be misclassified.

Example 2:

For a cutting score of .85, suppose a practitioner is willing to accept a 10% misclassification rate for students whose domain score is .95. How many questions should the random sample (test) have, and what should the advancement score be?

Answer: 11 items with an advancement score of 10.

The primary problem in applying the tables prepared by Millman (1973) is that one would need to have a good prior estimate of an examinee's domain score. Other problems have been suggested by Novick and Lewis (1974): They reported that for certain combinations of cut-off scores and test length, changing one or both to decrease the probability of misclassification for those above the cut-off score will actually increase the probability of misclassification for those below the cut-off score. In order to choose the appropriate combination of test length and advancement score, one must have some idea of whether the preponderance of examinees are above or below the cut-off score and one must have some idea of the relative costs of misclassification. However, the first requirement can only be satisfied with prior information about the domain scores of the group of examinees. Novick and Lewis (1974) suggested that it would be useful to have some systematic way of incorporating prior knowledge into the test length determination problem.

Table 4.2.2 below, from Novick and Lewis (1974), highlights the problem they raise:

Table 4.2.2 Percent of Students Expected To Be Incorrectly Advanced or Retained

Criterion Level = .75 Test Length = 8

Advancement	Domain Score Level									
Score	50	55	60	65	70	75	80	85	90	95
6	15	22	32	43	55	32	20	11	4	1
7	4	7	11	11	26	63	50	34	19	6

Suppose 7 out of 8 were taken as the minimum advancement score. Then for students whose true levels are <.75, the probabilities of misclassification fall off dramatically, while for students whose true levels are >.75 (more than likely where most of the students are located), these probabilities remain quite high. This would be the area where one would want the probability to be lower. Novick and Lewis conclude that a "framework would need to take into account on which side of .75 small expected errors were considered to be more important."

*4.2.4 Novick and Lewis' Bayesian Approach: Introduction[1]

Instead of considering the probability that a student will attain a test score, given his/her true level (an unknown), it would be better to consider the probability that a student's domain score exceeds a given cutting score, given his/her test score. A student will then be passed on to the next unit only if there is a sufficiently high probability that his/her domain score exceeds the cutting score, given his/her test score. The procedures offered by Novick and Lewis allow such a probability to be assessed. According to Novick and Lewis:

---

[1] An excellent introduction to Bayesian methods is given by Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974. Chapter 5 is especially relevant for our work in this unit.

> To obtain the necessary probability an application of Baye's theorem is required. In such an analysis prior knowledge (expressed in probabilistic terms) of the student's true level of functioning is combined with the (binomial) model information relating the observed test score to true level: the result is a posterior probability for true level of functioning, given the test score. The probability this distribution assigns to levels above the criterion is the quantity of interest.

Novick and Lewis produced a table (Table 4.2.3) reporting values of Prob $(\pi \geq \pi_o | x, n)$, i.e., the probability of an examinee having a domain score greater than or equal to a cut-off score π_o with a proportion correct score of x/n, for typical values of π_o, x, and n used in objectives-based instructional programs. (Actually the test lengths considered in their paper are a little longer than those often used in practice. The shortness of many criterion-referenced tests in use today is due in part to the failure of users to have any idea about the number of classification errors that are made with criterion-referenced tests.) In Table 4.2.3 π_o takes on values ranging from .50 to .95 (in increments of .05), test scores vary from 6 to 11, and test lengths vary from 8 to 12. Their table can be used to select both an advancement score and test length to ensure that Prob $(\pi \geq \pi_o)$ is larger than some desired value (say 70%). For example, if an instructor desired to ensure that Prob $(\pi \geq .80)$ was greater than .70, using the Novick-Lewis Table 4.2.3, it can be seen that an examinee should achieve 8 of 8 test items.

Table 4.2.3 Probability Student's Domain Score Is Greater Than π_o Given a Uniform Prior Distribution

Minimum Advancement Score	No. of Test Items	Posterior Distribution	Cut-off Score—π_o									
			50	55	60	65	70	75	80	85	90	95
6	8	β(7,3)	91	85	77	66	54	40	26	14	5	1
7	8	β(8,2)	98	96	93	88	80	70	56	40	23	7
8	8	β(9,1)	100	100	99	98	96	92	87	77	61	37
7	9	β(8,3)	95	90	83	74	62	47	32	18	7	1
8	9	β(9,2)	99	98	95	91	85	76	62	46	26	9
9	9	β(10,1)	100	100	99	99	97	94	89	80	65	40
7	10	β(8,4)	89	81	70	57	43	29	16	7	2	-
8	10	β(9,3)	97	93	88	80	69	54	38	22	9	2
9	10	β(10,2)	99	99	97	94	89	80	68	51	30	10
8	11	β(9,4)	93	87	77	65	51	35	21	9	3	-
9	11	β(10,3)	98	96	92	85	75	61	44	26	11	2
10	11	β(11,2)	100	99	98	96	92	84	73	56	34	12
9	12	β(10,4)	95	91	83	72	58	42	25	12	3	-
10	12	β(11,3)	99	97	94	89	80	67	50	31	13	2
11	12	β(12,2)	100	100	99	97	94	87	77	60	38	14

(Tables 4.2.3 thru 4.2.11 are reproduced from Novick and Lewis, 1974, with permission from the authors.)

A. Examples of the Effects of Different Priors

In this section, we will present tables from Novick and Lewis that demonstrate the effects of specifying different priors. Consider two situations. We will label them "A" and "B".

In situation A, we know very little at all about a student's domain score prior to test administration. Hence, we select as our prior a uniform distribution ($\beta[1,1]$) on the interval from zero to unity. Table 4.2.3 provides the posterior probabilities for various test lengths and cut-off scores.

To use Table 4.2.3 to select test length, one must decide on the cut-off score and the minimum acceptable probability that a student's domain score exceeds this cut-off score.

Example 3:

If we take the cut-off score (π_o) to be .80 and the minimum probability to be .5 (e.g., Prob ($\pi \geq \pi_o | x,n$) = .5 where x = test score, n is test length), what is the minimum number of test items that can be used, and what is the minimum advancement score?

Answer: 8 items with an advancement score of 7, because Prob ($\pi \geq \pi_o | 7,8$) = .56, which is greater than .50.

Example 4:

Suppose the cut-off score is π_o = .90 and we want Prob ($\pi \geq .9 | x,n$) = .5, what is the minimum number of test items that can be used, and what is the minimum advancement score?

Answer: There is no number of test items and advancement score less than perfection that satisfies this condition for the items specifications in the table. Note that Prob $(\pi \geq .9|8,8) = .61$ and Prob $(\pi \geq .9|9,9) = .65$. The answer is 8 items, and an advancement score of 100% (8 out of 8 items answered correctly).

In situation B, suppose that our probability that a student is functioning above the cutting score of .80 is .75. (When we specified the uniform prior, we set the prior probability at .20.) Novick and Lewis note that this belief can be characterized by the beta prior distribution $\beta(10.254, 1.746)$. Table 4.2.4 gives the same sort of information as Table 4.2.3, but is based upon a revised prior.

To use Table 4.2.4 one must again set a cut-off score and decide on a minimum acceptable probability that a student's domain score exceeds this cutting score.

Example 5:

Suppose the cut-off score is $\pi_o = .90$ and we want Prob$(\pi \geq .9|x,n) = .5$. What is the minimum number of test items that can be used, and what is the minimum advancement score? [Assume the prior is given by $\beta(10.254, 1.746)$.]

Answer: For 12 items with an advancement score of 11, we have Prob$(\pi \geq .9|11,12) = .48$, which is sufficiently close to .50. (Shorter test lengths can be chosen—8 and 9 test items--if the advancement score is set at 100%.)

Table 4.2.4. Probability Student's Domain Score Is Greater Than π_0 Given a $\beta(10.254, 1.746)$ Prior Distribution

Minimum Advancement Score	No. of Test Items	Posterior Distribution	Criterion Level—π_0 50	55	60	65	70	75	80	85	90	95
6	8	$\beta(16.254, 3.746)$	100	100	98	96	90	78	60	37	15	2
7	8	$\beta(17.254, 2.746)$	100	100	100	99	97	92	81	62	36	10
8	8	$\beta(18.254, 1.746)$	100	100	100	100	99	98	94	85	66	32
7	9	$\beta(17.254, 3.746)$	100	100	99	97	92	82	65	41	17	2
8	9	$\beta(18.254, 2.746)$	100	100	100	99	98	93	84	66	39	11
9	9	$\beta(19.254, 1.746)$	100	100	100	100	100	98	95	87	69	34
7	10	$\beta(17.254, 4.746)$	100	99	97	93	84	68	47	24	7	1
8	10	$\beta(18.254, 3.746)$	100	100	99	98	93	84	68	45	19	3
9	10	$\beta(19.254, 2.746)$	100	100	100	99	98	95	86	69	42	12
8	11	$\beta(18.254, 4.746)$	100	99	98	94	87	72	51	27	8	1
9	11	$\beta(19.254, 3.746)$	100	100	100	98	95	87	72	48	22	3
10	11	$\beta(20.254, 2.746)$	100	100	100	100	99	96	88	72	45	13
9	12	$\beta(19.254, 4.746)$	100	100	99	96	89	76	55	30	10	1
10	12	$\beta(20.254, 3.746)$	100	100	100	99	96	89	75	52	24	4
11	12	$\beta(21.254, 2.746)$	100	100	100	100	99	96	90	75	48	14

Note: The mean and mode, respectively of $\beta(10.254, 1.746)$ are .855 and .925 and for this distribution Prob $(\pi > \pi_0)$ for π_0 = .70, .75, .80, .85 are .92, .86, .75, and .59, respectively. A close look at these distributional characteristics will help a decision maker determine if this prior distribution is a realistic characterization of his/her beliefs.

(Taken from Novick and Lewis, 1974.)

We can assess the effects of prior information by looking at some representative situations and the probabilities associated.

Situation	Uniform Prior	$\beta(10.254, 1.746)$ Prior
$\text{Prob}(\pi \geq .8 \mid 6,8)$	.26	.60
$\text{Prob}(\pi \geq .8 \mid 10,12)$	.50	.75
$\text{Prob}(\pi \geq .9 \mid 6,8)$	.05	.15
$\text{Prob}(\pi \geq .9 \mid 10,12)$	.13	.24
$\text{Prob}(\pi \geq .7 \mid 7,9)$	.62	.92
$\text{Prob}(\pi \geq .7 \mid 9,11)$	.75	.95

These situations are provided as examples, but the message is clear; specifying a prior of the sort in situation B results in a much higher probability statement about an individual's domain score exceeding the cutting score, given the test data. According to Novick and Lewis:

> When the decision maker specifies an informative prior distribution he is saying, in effect, that he wants a decision which will have a high probability of being correct in that portion of the decision space in which he thinks the student's ability truly lies.

B. Losses Associated with Incorrect Decisions

Before discussing Novick and Lewis' tables for test length, we must discuss how they specify loss ratios. This is critical for use of the tables. This is consonant with the formulations for the decision-theoretic approach to setting cut-off scores (see Unit 6), but we will discuss these procedures here for continuity. The following two-fold table of losses associated with decisions can be constructed:

		Domain Score	
		$\pi \geq \pi_0$	$\pi < \pi_0$
Decision	advance	o	a
	retain	b	o

Where π_0 = cutting score on domain of tasks

a = loss associated with advancing a student whose true level $\pi < \pi_0$ (false positive error)

b = loss associated with retaining a student whose true level $\pi \geq \pi_0$ (false negative error)

Suppose a= b. According to Novick and Lewis:

> If it were no more serious to advance a student whose level was below the criterion than to retain a student who was above, we would be behaving optimally if we were to advance students with posterior probabilities above .5 and retain the others.

Novick and Lewis further point out that if the loss for false advancement is twice that of false retention, which is a more reasonable situation, then only those students whose posterior probabilities are greater than $\frac{2}{3} = .67$ should be advanced.

More generally, the decision rule to be used is to advance a student if his/her test score is such that $(b)[\text{Prob}(\pi \geq \pi_0 | x,n)] \geq (a)[\text{Prob}(\pi < \pi_0 | x,n)]$ and retain him/her if this is not true. An equivalent procedure is to compare the loss ratio $\frac{a}{b}$ (It would be $\frac{2}{1}$ above) to the ratio

$$\frac{\text{Prob}(\pi \geq \pi_0 | x,n)}{\text{Prob}(\pi < \pi_0 | x,n)}.$$

Various loss ratios are specified in the tables to be discussed next.

C. Test Length Specifications

In order to use the tables that follow, one must specify the following:

1. A criterion level, or cutting score, π_0, must be chosen.
2. Prior knowledge of student's domain score must be translated into a prior probability distribution of the β form for π (Use the methods described in Novick and Jackson, 1974).
3. A loss ratio $\frac{a}{b}$ for the relative losses associated with the two types of incorrect decisions must be chosen.

From these specifications, the tables give:

1. recommended test lengths,
2. minimum advancement scores,
3. posterior probability that the domain score is greater than π_0, given the test data, and
4. the percentage correct specified by the advancement rule for the recommended sample size(s).

Table 4.2.5 provides some beta distributions and corresponding parameters; it will be helpful to individuals setting priors.

Before providing some examples, we should comment on Tables 4.2.8 and 4.2.9, which appear to be the same. If carefully scrutinized, one will notice that the expected values of the prior distributions are different, and this changes the entries in the body of the table. In Table 4.2.8, the expected value of the prior is .8, which equals π_0, while in Table 4.2.9, it is .85, which is larger than π_0. The sample sizes that are recommended in Table 10 are clearly more attractive. For instance, for a beta prior $\beta(12,3)$ and the expected value = .8, the test would be 22 items with an advancement score of 19, while when the expected value = .85, the recommended test drops to only 13 items with a passing score of 11. Novick and Lewis comment as follows:

> When loss ratios are high it may well be advantageous to strengthen the training program to the extent that the the mean output is well above the specified criterion level. This will make it possible to use short tests, or, alternatively, will generally reduce the risk of incorrect classification.

Example 6:

You have decided on a cutting score $\pi_0 = .8$ and your prior has been computed to be $\beta(8,2)$. You have decided on a loss ratio of 2.5 (it is 2.5 times as costly to advance someone whose $\pi<\pi_0$ than to retain someone whose $\pi\geq\pi_0$). What is the recommended test length and advancement score, and also the associated probability?

Answer: 20 questions with an advancement score of 17 (re: 85% mastery based on test). $\text{Prob}(\pi\geq.8|17,20) = .72$. We are 72% sure an individual's domain score is above .80.

Table 4.2.5 Selected Prior Distributions for Advancement Decisions

No.	Prior Distribution	Effective Prior Sample Size	Mean	Prob ($\pi_l \le \pi \le \pi_u$)* .00-.70	.70-.75	.75-.80	.80-.85	.85-.90	.90-1.00
1	β(5.6, 2.4)	8	.70	.46	.12	.12	.12	.10	.08
2	β(6,2)	8	.75	.33	.12	.13	.14	.13	.15
3	β(6.4, 1.6)	8	.80	.21	.10	.12	.15	.16	.26
4	β(6.8, 1.2)	8	.85	.12	.07	.09	.13	.17	.42
5	β(7.2, .8)	8	.90	.05	.04	.06	.09	.14	.62
6	β(7, 3)	10	.70	.46	.14	.14	.12	.09	.05
7	β(7.5, 2.5)	10	.75	.32	.13	.15	.15	.13	.12
8	β(8, 2)	10	.80	.20	.10	.14	.16	.17	.23
9	β(8.5, 1.5)	10	.85	.10	.07	.10	.14	.19	.40
10	β(9, 1)	10	.90	.04	.03	.06	.10	.16	.61
11	β(8.4, 3.6)	12	.70	.47	.15	.15	.12	.08	.03
12	β(9, 3)	12	.75	.32	.14	.16	.16	.13	.09
13	β(9.6, 2.4)	12	.80	.18	.11	.15	.18	.18	.20
14	β(10.2, 1.8)	12	.85	.09	.07	.11	.16	.20	.37
15	β(10.8, 1.2)	12	.90	.03	.03	.06	.11	.17	.60
16	β(10.5, 4.5)	15	.70	.47	.17	.16	.12	.06	.02
17	β(11.25, 3.75)	15	.75	.30	.16	.18	.17	.13	.06
18	β(12.3)	15	.80	.16	.12	.17	.20	.19	.16
19	β(12.75, 2.25)	15	.85	.07	.07	.12	.18	.23	.33
20	β(13.5, 1.5)	15	.90	.02	.03	.06	.11	.19	.59

Note: All entries have been rounded to two decimal places and smoothed so that the row totals add to 1.00.

Table 4.2.6 Recommended Sample Sizes and Advancement Scores

Prior Distribution	$\varepsilon(\pi)$	$\pi_0 = .70$ Loss Ratio 1.5(.60)	2.0(.67)	2.5(.71)	3.0(.75)
$\beta(5.6, 2.4)$[1]	(.70)	6/8(.62)	10/13(.70)	11/14(.74)	12/15(.78)
$\beta(7, 3)$	(.70)	6/8(.61)	10/13(.69)	11/14(.73)	12/15(.77)
$\beta(8.4, 3.6)$	(.70)	6/8(.61)	10/13(.68)	11/14(.72)	12/15(.76)
$\beta(10.5, 4.5)$	(.70)	9/12(.62)[2]	10/13(.67)	11/14(.71)	12/15(.75)
			General Recommendations		
		6/8(75%)	10/13(77%)	11/14(79%)	12/15(80%)

[1] Apriori, Prob($\pi \geq .70$) for each of the four prior distributions is .54, .54, .53, and .53.

[2] For 6/8, Prob($\pi \geq .70$) = .598.

Table 4.2.7 Recommended Sample Sizes and Advancement Scores

		$\pi_o = .75$			
		Loss Ratio			
Prior Distribution	ε(π)	1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
β(6, 2)[1]	(.75)	8/10(.65)	16/20(.70)	17/21(.74)	18/22(.77)
β(7.5, 2.5)	(.75)	8/10(.64)	16/20(.69)	17/21(.73)	18/22(.76)
β(9, 3)	(.75)	8/10(.63)	16/20(.69)	17/21(.72)	18/22(.75)
β(11.25, 3.75)	(.75)	8/10(.62)	16/20(.68)	17/21(.71)	19/23(.77)[2]
		General Recommendations			
		8/10(80%)	16/20(80%)	17/21(81%)	18/22(82%)

[1] Apriori, Prob(π≥.75) = .56, .55, .55, and .54, respectively for the four prior distributions used in Table 7.

[2] For 18/22, Prob(π≥.75) = .744.

Table 4.2.8 Recommended Sample Sizes and Advancement Scores

		$\pi_o = .80$			
		Loss Ratio			
Prior Distribution	ε(π)	1.5 (.60)	2.0 (.67)	2.5 (.71)	3.0 (.75)
β(.64, 1.6)[1]	(.80)	6/7(.66)	7/8(.70)	17/20(.72)	19/22(.78)
β(8, 2)	(.80)	6/7(.65)	7/8(.69)	17/20(.72)	19/22(.77)
β(9.6, 2.4)	(.80)	6/7(.64)	7/8(.68)	17/20(.71)	19/22(.76)
β(12, 3)	(.80)	6/7(.63)	7/8(.67)	18/21(.73)[2]	19/22(.75)
		General Recommendations			
		6/7(86%)	7/8(88%)	17/20(85%)	19/22(86%)

[1] Apriori, Prob(π≥.80) = .57; for 8/10, Prob(π≥.80) = .55; for 16/20, Prob(π≥.80) = .54; for 8.5/10, Prob(π≥.80) = .67; for 8.3/10, Prob(π≥.80) = .62; for 9/10, Prob(π≥.80) = .78.

[2] For 17/20, Prob(π≥.80) = .70.

Table 4.2.9 Recommended Sample Sizes and Advancement Scores

		$\pi_o = .80$			
		Loss Ratio			
Prior Distribution	$\varepsilon(\pi)$	1.5(.60)	2.0(.67)	2.5(.71)	3.0(.75)
$\beta(6.8, 1.2)$[5]	(.85)	8/10(.64)	9/11(.69)	10/12(.72)[1]	11/13(.76)
$\beta(8.5, 1.5)$	(.85)	8/10(.66)	9/11(.70)	10/12(.73)[2]	11/13(.76)
$\beta(10.2, 1.8)$	(.85)	8/10(.67)	9/11(.71)	9/11(.71)[3]	11/13(.77)
$\beta(12.75, 2.25)$	(.85)	8/10(.69)	9/11(.72)	9/11(.72)[4]	11/13(.78)
		General Recommendations			
		8/10(80%)	9/11(82%)	10/12(83%)	11/13(85%)

[1] For 5/6, Prob($\pi \geq .80$) = .72.

[2] For 5/6, Prob($\pi \geq .80$) = .73.

[3] For 10/12, Prob ($\pi \geq .80$) = .74.

[4] For 10/12, Prob ($\pi \geq .80$) = .75.

[5] For the four prior distributions, the apriori probabilities of $\pi \geq .80$ are .72, .73, .74, and .75. With these prior distributions and with 7/10, the posterior probabilities of $\pi \geq .80$ are .41, .43, .46, and .48.

Table 4.2.10 Recommended Sample Sizes and Advancement Scores

		$\pi_0 = .85$			
		Loss Ratio			
Prior Distributions	$\varepsilon(\pi)$	1.5 (.60)	2.0 (.67)	2.5 (.70)	3.0 (.75)
$\beta(6.8, 1.2)$[1]	(.85)	7/8(.62)	9/10(.70)	17/19(.73)	18/20(.76)[3]
$\beta(8.5, 1.5)$	(.85)	7/8(.62)	9/10(.69)	17/19(.72)	19/21(.77)
$\beta(10.2, 1.8)$	(.85)	7/8(.61)	9/10(.68)	17/19(.72)	19/21(.76)
$\beta(12.75, 2.25)$	(.85)	7/8(.60)	9/10(.67)	17/19(.71)[2]	19/21(.75)
		General Recommendations			
		7/8(87.5%)	9/10(90%)	17/19(89%)	19/21(90%)

[1] The apriori probabilities for $\pi > .85$ are .59, .58, .58, and .57.

[2] For 10/11, Prob($\pi > .85 = .695$).

[3] For 19/21, Prob($\pi > .85 = .78$).

Example 7:

You have decided on a cutting score $\pi_o = .85$ and your prior is $\beta(12.75, 2.25)$. Your loss ratio is 1.5. What is the recommended test length, advancement score, and associated probability?

Answer: 8 questions with an advancement score of 7 (re: 87.5%) $\text{Prob}(\pi > .85 \mid 7,8) = .60$; we are 60% sure an individual's domain score is above .85.

D. Suggestions

We suggest that the tables developed by Novick and Lewis be used any time you can specify a meaningful prior. As mentioned before, the tables developed by Millman are meaningful only for quick estimates. They give probabilities of test data, given domain score, what is really needed is the probability of domain score, given test data. We recommend that if there is no suitable prior, Table 4.2.1 in this section be consulted. Also, for such a situation, the methods developed by Fhaner and Wilcox, involving the specification of an indifference zone, should be considered.

4.2.5 Fhaner (1974) and Wilcox (1976) Use of Indifference Zones

What follows is a discussion of the use of indifference zones merging the work of Fhaner and Wilcox, using Wilcox's notation. The basic situation is that described in section 1, and is similar to the section on Millman's procedures.

The binomial distribution can be used to estimate the probability of an examinee whose domain score is π obtaining a test score of x items out of n items.

$$\text{Prob } (x|\pi) = \binom{n}{x} \pi^{x} (1-\pi)^{n-x}$$

where $\frac{x}{n}$ is an unbiased estimate of π.

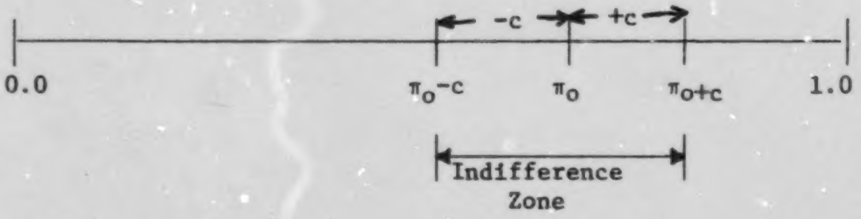
Tests are used in a context; the context for criterion-referenced testing in decision making, where the test score will be used to classify individuals. To separate individuals into mastery states (mastery versus non-mastery), a cutting score π_0 is established such that if $\pi<\pi_0$ the examinee is a non-master; if $\pi\geqslant\pi_0$ the examinee is a master. The tester has only the test score x to work with, not π, and needs to decide if $\pi<\pi_0$ or $\pi\geqslant\pi_0$. Hence, there is the risk of false positive errors ($\pi<\pi_0$, but the examinee passes based on the test) or false negative errors ($\pi\geqslant\pi_0$, but the examinee fails based on the test). Let α be the probability of a Type I (false positive error) and β be the probability of a Type II (false negative error). A performance score n_0 needs to be established such that:

$$\text{Prob}(x \geqslant n_0|\pi) \leq \alpha \text{ for all } \pi<\pi_0$$

$$\text{Prob}(x < n_0|\pi) \leq \beta \text{ for all } \pi\geqslant\pi_0$$

Since $\alpha = 1-\beta$, it is not possible to keep both probabilities at acceptably low levels. An explicit solution to the problem is generated by establishing an indifference zone. Let c be a positive constant, and form the open interval $(\pi_0 - c, \pi_0 + c)$. For individuals whose domain score is close to π_0 (within the interval from $\pi_0 - c$ to $\pi_0 + c$), we are "indifferent" as to how they are classified, re: there is negligible loss in misclassification of such individuals. For individuals whose domain score is greater than $\pi_o + c$ or less than $\pi_o - c$, we want to be reasonably certain the correct decision is made. Schematically,

Domain Score Scale



Thus far we have been working with the domain of tasks. We must now specify procedures involving the test itself. Let n_o = passing score or advancement score on the test. Thus, if $x \geq n_o$, the student is advanced; if $x < n_o$, the student is retained. A correct decision is made for the student if $x < n_o$ and $\pi < \pi_o$ or $x \geq n_o$ and $\pi \geq \pi_o$. Let P^* be a number such that $\frac{1}{2} < P^* < 1$. Our goal is to establish n as small as possible (for a certain n_o) so that for values of π not in the indifference zone, the probability of a correct decision is at least P^*.

For values of $\pi \leq \pi_o - c$, the minimum probability of a correct decision occurs at the point $\pi_o - c$ and is given by

$$\alpha = \sum_{x=o}^{n_o-1} \binom{n}{x} (\pi_o - c)^x (1 - \pi_o + c)^{n-x}$$

For values of $\pi \geq \pi_o + c$, the minimum probability of a correct decision occurs at the point $\pi_o + c$ and is given by:

$$\beta = \sum_{x=n_o}^{n} \binom{n}{x} (\pi_o + c)^x (1 - \pi_o - c)^{n-x}$$

Now to choose n, Wilcox specifies:

> In particular, we choose the smallest integer n so that α and β are greater than or equal to P* which implies that the probability of a correct decision is at least P* for $\pi \geq \pi_0 + c$ and $\pi \leq \pi_0 - c$.

Wilcox provides tables for various combinations of the variables involved in the formula. In order to use these tables, the following must be specified:

1. π_0: The cutting score for the domain of items. Wilcox specifies the π_0's to be .70, .75, .80, .85..
2. c: The positive constant that forms the indifference zone. Wilcox uses c = .05 and c = .10. Thus, for π_0 = .75 and c = .10, we are indifferent as to our classification for scores in the interval (.65, .85).
3. P*: The minimum probability of a correct decision for scores not in the indifference region. Wilcox uses P* = .75.

By specifying these values, Wilcox's table then gives you n and n_0, along with the probability of correctly classifying an examinee with a domain score $\geq \pi_0 + c$ or $\leq \pi_0 - c$.

Example 8:

Suppose π_0, the cut-off score = .80, c = .1 and P* = .75. What is the least number of questions that can be used to have greater than a 75% chance of correctly classifying an examinee in the interval greater than $\pi_0 + c$ (= .9) and less than $\pi_0 - c$ (= .7).

Answer: For π_0 = .8, c = .1, the least number of questions n is 9 with an advancement score of 8.

Table 4.2.11

Cut-off Scores and the Minimum Probability of a Correct Decision for Values of π not in the Indifference Zone

n	$\pi_o = .70$		$\pi_o = .75$		$\pi_o = .80$		$\pi_o = .85$	
	c=.05	c=.10	c=.05	c=.10	c=.05	c=.10	c=.05	c=.10
8	6/.5722	6/.6846	7/.5033	7/.6572	7/.6329	7/.7347	7/.4967	7/.6329
9	7/.6007	7/.7382	7/.5372	7/.6627	8/.5995	8/.7748	8/.5683	8/.6997
10	8/.5256	8/.6778	8/.6172	8/.7384	9/.5443	9/.7361	9/.6242	9/.7560
11	8/.5744	8/.7037	9/.6174	9/.7788	9/.5448	10/.6974	10/.6779	10/.8029
12	9/.6488	9/.7747	10/.5583	10/.7358	10/.6093	10/.7472	11/.6590	11/.8416
13	10/.5843	10/.7473	10/.5794	11/.7296	11/.6674	11/.7975	12/.6213	12/.8646
14	10/.5733	10/.7207	11/.6488	11/.7795	12/.6479	12/.8392	13/.5846	13/.8470
15	11/.6481	11/.7827	12/.6482	12/.8227	13/.6042	13/.8159	13/.6020	14/.8290
16	12/.6302	12/.7982	13/.5981	13/.7899	13/.5950	14/.7892	14/.6482	15/.8108
17	12/.5803	13/.7582	13/.6113	13/.7652	14/.6470	14/.7981	15/.6904	15/.8363
18	13/.6450	13/.7912	14/.6673	14/.8114	15/.6943	15/.8354	16/.7287	16/.8647
19	14/.6678	14/.8369	15/.6733	15/.8500	16/.6841	16/.8668	17/.7054	17/.8887
20	14/.6172	15/.8042	16/.6296	16/.8298	17/.6477	17/.8670	18/.6769	18/.9087

(Reproduced from Wilcox, 1976, permission pending.)

Example 9:

Suppose $\pi_0 = .7$ and $c = .05$. On a 15 item test with an advancement score of 11, what is the probability of correctly classifying an individual with a domain score $\geq .75$ ($= \pi_0 + c$) or $\leq .65$ ($= \pi_0 - c$).

Answer: 65%

A. Summary of the Procedure

Because of the complexity of this formulation, a summary would appear helpful. One wants to minimize the probabilities of incorrectly classifying individuals based upon a test score, i.e., one wants to minimize the probability of a false positive error (α) and the probability of a false negative error (β) simultaneously. Since $\alpha = 1 - \beta$, minimizing one will maximize the other. The problem is circumvented by specifying an indifference zone, and then α and β can be simultaneously minimized at the boundary points of the indifference zone. Wilcox has prepared tables that relate the number of test items to α and β, for specific indifference zones, in terms of probability of correct decisions outside the indifference zone.

B. Comments

Consider next the work of Fhaner and Wilcox:

1. If $c = 0$, that is, there is no indifference region, it is not always possible to choose n such that the probability of a correct decision is at least P*. Wilcox says that for this situation the probability of a correct decision approaches .5 (an unacceptable level) as n increases. Hence, Millman's solution may not be adequate for certain situations

2. If the loss in misclassifying an individual who has obtained mastery ($\pi \geq \pi_o + c$) is different from the loss in misclassifying a non-master ($\pi \leq \pi_o - c$), then two numbers P_1^* and P_2^* can be chosen such that $\frac{1}{2} < P_1^* < 1$ and $\frac{1}{2} < P_2^* < 1$ and there is a smallest integer n so that $\alpha \geq P_1^*$ and $\beta \geq P_2^*$.

3. If n is large, a theorem in statistics called the Central Limit Theorem justifies the use of the normal distribution in place of the binomial. In this case, tables of the normal distribution function (Φ) may be used, and use of the Wilcox tables can be circumvented. In this case, the number of test questions is given by:

$$n = \left(\frac{Z_{1-\alpha}\sqrt{(\pi_o - c)(1-\pi_o + c)} + Z_{1-\beta}\sqrt{(\pi_o + c)(1 - \pi_o - c)}}{2c} \right)^2$$

where n = number of items

c = positive constant (from before)

π_o = cutting score

$Z_{1-\alpha}$ = deviation score in a standardized normal distribution corresponding to $1-\alpha$

$Z_{1-\beta}$ = deviation score in a standardized normal distribution corresponding to $1-\beta$.

Fhaner notes that the normal approximation underestimates the number of items needed. Wilcox notes that the procedure does not give you an optimal n_o (performance or advancement score). Hence, a user needs to be careful when making use of the normal approximation.

4.2.6 Eignor-Hambleton Approach for Determining Test Length[1]

Methods for determining test length which depend upon the minimization of classification errors were presented in the last three sections. An alternate approach in which test length is related directly to several indices of test score reliability and validity is presented in this section. At this point it would be desirable for the reader to study material presented in Unit 5 on test score reliability before proceeding further.

A. Introduction

A primary concern of individuals using test scores is that the scores be both reliable and valid. While the best approach for assessing test score reliability and validity will depend on the particular situation, it is well-known that there is a relationship between the length of a test, the advancement score, and the reliability and validity of the test scores. Longer tests result in test scores with better psychometric properties.

For norm-referenced tests, the relationship of test length to reliability can be expressed by the Spearman-Brown formula. Also, formulas exist that relate norm-referenced test length to test score validity. However, because these formulas are based upon a correlational approach to reliability and validity, they are not very useful with criterion-referenced tests when the intent of the criterion-referenced test is to produce scores for making mastery/non-mastery decisions (Hambleton, Swaminathan, Algina, & Coulson, 1978). What is often of interest to users of criterion-referenced tests is information concerning the consistency of mastery/non-mastery decisions for some group of examinees across a retest administration or across a parallel-form administration. Second, there is usually considerable interest in the extent of agreement between mastery/non-mastery decisions based on a criterion-referenced test and the "true" mastery states of a group of examinees (sometimes called "decision accuracy"). (The "true" mastery state of an examinee is the one he/she should be assigned to, based on the amount of knowledge or skill he/she possesses relative to the objective or competency under investigation.) These two situations described above correspond to one paradigm for viewing the psychometric concepts of criterion-referenced test score reliability and validity, respectively.

---

[1] Material in this section is from a paper by Eignor and Hambleton (1979). Additional results are reported in Eignor (1979).

Hambleton et al. (1978) distinguished between uses of criterion-referenced test scores, domain score estimation and allocation of examinees to mastery states. For the first use, the test length relationship to reliability can be derived, and may be summarized in the well-known item sampling model (Lord and Novick, 1968). It is for the other major use of criterion-referenced test scores, mastery state determination, that necessary technical developments are in short supply. Little research has been done that directly explores the relationships of test length and cut-off scores to criterion-referenced test score reliability and validity when the scores are used for assigning examinees to mastery states.

What research has been done has focused either (1) on procedures for determining reliability of examinee assignments to mastery states (Hambleton & Novick, 1973; Swaminathan, Hambleton, & Algina, 1974; Huynh, 1976; Subkoviak, 1976, 1978a, 1978b; Marshall & Haertel, 1976; Algina & Noe, 1978) or (2) on procedures for the determination of test length that minimizes misclassification errors (Millman, 1973; Novick & Lewis, 1974; Fhaner, 1974; Wilcox, 1976, 1977). The research reported in this paper is directed toward linking together these two areas of research and providing useful results to test practitioners to enable them to determine test lengths to fit the situations in which their tests will be used.

Specifically, the purpose of the study was two-fold:

1. To report the relationships between test lengths and several reliability and validity indices for a fixed cut-off score (80%) in five domain score distributions.
2. To report the relationships between advancement scores and several reliability and validity indices for several test lengths and in five domain score distributions.

The study was carried out using computer simulation methods. The one major advantage of this approach is that it is possible "to know" examinee domain scores and their "true" mastery states. Such information permits one to compare examinee estimated domain scores and assigned mastery states based on test results with domain scores and true mastery states. Summary of such comparisons address the validity of the particular set of test scores under investigation.

B. Research Design

Terminology

Test length refers to the number of test items that are used to measure examinee performance on a particular objective. A domain score for an examinee is defined as the proportion of items in the domain of items measuring an objective that the individual can answer correctly. A cut-off score is set on the domain score scale [0,1] to separate examinees into two true mastery states.

Since all items in the domain of items defined by an objective cannot usually be administered to examinees for the purpose of assessing their domain scores or assigning them to mastery states, a sample of test items is chosen. Estimated domain score is defined as the proportion of items that an examinee answers correctly of the items included in the test. An advancement score is defined as the number of items on the test measuring an objective deemed necessary for an individual to answer correctly to be classified as a master.

In using an examinee's test score to determine his/her true mastery status, two types of classification errors can result. A false-positive error occurs when an examinee is estimated to be a master when his/her true status is non-master; a false-negative error occurs when an examinee is estimated to be a non-master when his/her true status is master.

Variables Under Study

(a) Test Model

Both the binomial and compound binomial models were used to simulate examinee item response data. While criterion-referenced test data has often been assumed to fit the binomial model, Lord (1965), and more recently, Wilcox (1976, 1977), have suggested that the compound binomial model may be appropriate. The binomial model assumes that the probability of a correct response for an examinee is the same across all items on a test; or alternatively, that all items are equally difficult (for that examinee). The compound binomial model assumes that the probability of correct response for an examinee varies across items in a test, or that the items are not equally difficult (for that examinee). The latter assumption is considerably more plausible but investigations that have utilized both models (for instance, Subkoviak, 1976) have demonstrated different, but not very much different results from the use of the two models.

(b) Prior Distributions

For the binomial model, either a user-supplied or a beta prior distribution on domain scores was specified and examinee domain scores sampled from this distribution. For the user-supplied prior, a percentage of examinees were assigned to each of ten equal intervals from 0.00 to 1.00, and a domain score distribution was constructed from this information. The percentages assigned to the intervals reflect the prior belief about how the group would perform on the domain of tasks of which the test is a sample. An examinee's domain score was then sampled from this prior distribution, and this domain score used to simulate binomial model test performance. This process was then repeated for 200 examinees.

When the prior domain score distribution was specified as a beta prior, the fractile assessment procedure (Novick & Jackson, 1974) was used to specify the parameters of the beta distribution, and then a IMSL Subroutine (GGBTA) used to generate the distribution. The justification for using a beta prior distribution stems from two facts. One, the beta distribution is defined on the interval [0,1] (whereas most other distributions are not). Second, the beta distribution allows the user to easily generate skewed domain score distributions to approximate distributions that might be expected to occur with real criterion-referenced test data.

The fractile assessment procedure (FASP) has been offered by Novick and Jackson (1974) as a means for specifying the parameters of a beta distribution. The user is asked to specify q_1, q_2, and q_3, the first, second (median), and third quartiles of the distribution. The parameters, a and b, of the beta distribution are then (approximately) given by:

$$a = cq_2 + \frac{1}{3} \text{ and } b = c(1-q_2) + \frac{1}{3}$$

where

$$c = .057\ (\frac{1}{d_1} + \frac{1}{d_3})$$

where

$$d_1 = \left[[q_2\ (1-q_1)]^{\frac{1}{2}} - [q_1\ (1-q_2)]^{\frac{1}{2}} \right]^2$$

and

$$d_3 = \left[[q_2\ (1-q_3)]^{\frac{1}{2}} - [q_3\ (1-q_2)]^{\frac{1}{2}} \right]^2 .$$

The parameters a and b were then used as input to the GGBTA Subroutine, which generated beta-distributed domain scores. As with the "user-supplied" prior distribution, domain scores were then used to simulate examinee binomial model test performance.

Domain scores for use with the compound binomial model were generated from a normal distribution (mean = 1, standard deviation = 1) and then rescaled to the interval [0,1]. This step (and others done with the compound binomial model) was carried out with the aid of computer program DATAGEN (Hambleton and Rovinelli, 1973). In the past it has been regularly used to generate logistic test model data.

Additional details on the five domain score distributions used in the study are reported in Table 1.

(c) Advancement Scores

In addressing the first purpose of the study, advancement scores were always set exactly equal to the chosen cut-off score of .80. This was made possible because of the test lengths under consideration.

In the second part of the study, for several fixed test lengths, advancement scores were moved around with the same test data sets, to determine the influence of advancement score placement on indices of test score reliability and validity.

(d) Test Lengths

Test lengths of 5, 10, 15, 20, and 40 were considered in this particular study. Many other test lengths (and advancement scores) were considered by Eignor (1979).

Table 1

Descriptions of the Five Domain Score Distributions

Distribution	Test Model	Skewness	Domain Score Distribution Description
1	Binomial	Moderate Negative	(a) Mode is slightly below the cut-off score (.80). (b) Range of scores is [.11, 1.00]. (c) About 50% are on the interval [.60, .80] and 80% on the interval [.50 to .90].
2	Binomial	High Negative	(a) Leptokurtic distribution with the mode above the cut-off score (.80). (b) Range of scores is [.60, 1.00] with about 80% of the scores on the interval [.80, 1.00].
3	Binomial	Slight Negative	(a) Mode is far below the cut-off score. (b) 50% on the interval [.00, .49] and 50% on the interval [.50, 1.00]. (c) Substantial variation of scores.
4	Compound Binomial	Moderate Negative	(a) Mode is close to the cut-off score. (b) Wide range of domain scores [.00, 1.00]. (c) 50% on the interval [.00, .79] and 50% on the interval [.80, 1.00]. (d) Flatter distribution than either (1) or (2).
5	Compound Binomial	None	(a) An almost rectangular distribution on the interval [.20, .90] with domain scores but fewer of them below .20 and above .90.

Reliability and Validity Indices

A number of practical indices of test score reliability and validity were used in the study. The two diagrams below will facilitate their discussion.

Diagram One

Test Occasion Two

		NM	M
Test Occasion One	NM	p_{00}	p_{01}
	M	p_{10}	p_{11}

Diagram Two

Criterion Measure

		NM	M
Test Results	NM	p_{00}	p_{01}
	M	p_{10}	p_{11}

(M = Mastery status; NM = Non-Mastery status)

The contingency table in diagram one shows the proportion of examinees falling in the four possible combinations of mastery state assignments based on parallel-form (or test-retest) administrations of a criterion-referenced test. The only difference in diagram two is that a criterion measure is substituted for a parallel-form of the criterion-referenced test under study.

Two reliability indices are derivable from data reported in Diagram One:

1. Decision Consistency

$$DC = \sum_{k=0}^{1} p_{kk} \qquad \text{(Hambleton \& Novick, 1973)}$$

2. Kappa

$$\kappa = \frac{DC-CA}{1-CA} \qquad \text{(Swaminathan, Hambleton, \& Algina, 1974)}$$

where CA (chance agreement) = $\sum_{k=0}^{1} p_{k\cdot}\, p_{\cdot k}$

and $p_{0\cdot}$, $p_{1\cdot}$, and $p_{\cdot 0}$, $p_{\cdot 1}$ are the respective marginal proportions for the first and second test administrations.

There are three derivable validity indices from Diagram Two:

3. Decision Accuracy

$$DA = \sum_{k=0}^{1} p_{kk} \qquad \text{(Hambleton and Novick, 1973)}$$

4. Predictive Validity (the Pearson correlation between decisions based on the criterion-referenced test and the criterion measure)

5. Efficiency

$$E = \frac{\sum_{i=1}^{N} (\pi_i - \pi_o) \text{ Sign } (\hat{\pi}_i - \hat{\pi}_o)}{\sum_{i=1}^{N} |\pi_i - \pi_o|} \qquad \text{(Livingston, 1978)}$$

where π_o is a cut-off score defined on the domain score scale, $\hat{\pi}_o$ is an advancement score, π_i is the domain score for examinee i and $\hat{\pi}_i$ is the estimated domain score for examinee i.

All of the statistics are well-known and commonly used in criterion-referenced testing practice except for the last one (and this is at least partially due to its newness). Essentially, efficiency is a measure of how accurately a criterion-referenced test and associated advancement score result in the assignment of examinees to mastery states that are in agreement with decisions based on a criterion measure. Also, the loss in efficiency due to misclassifying examinees (false-positive, and false-negative errors) is linearly related to the difference between an examinee's level of performance on the criterion test and the criterion test cut-off score. Clearly, Livingston's efficiency does not address directly the validity of mastery classifications. The index was included in the study because it provides an alternate but potentially useful framework for viewing criterion-referenced test score validity.

Data Generation

The process of generating examinee item scores and test scores and summary statistics on 200 examinees for various sets of testing conditions was completed as follows:

1. One of the domain score distributions from Table 1 and a test model (binomial or compound binomial) was selected.

2. Examinee domain scores were generated and examinees with domain scores equal to or above .80 were assigned to a mastery mastery state on the criterion measure. All other examinees were assigned to a non-mastery state.

3. For the particular test length under consideration, examinee domain score estimates were generated. For the binomial test model, this was done by setting the probability of a correct response for each item equal to the examinee's domain score. By generating random numbers uniformly distributed on the interval $[0, 1]$, it was possible to simulate the examinee's test item performance properly (i.e., answering each item with a probability of being correct equal to his/her domain score). Two sets of item scores were generated to simulate two test performances.

   For the compound binomial model, "item characteristic curves" were generated (see an example in Figure 1). From Figure 1 it is clearly seen that the probability of correct answers varies not only from one examinee to another but also for the same examinee from one item to another. Once probabilities for answering items for a given examinee are obtained, item scores via the use of a random number generator were obtained.

4. From the examinee item scores obtained in step 2, examinee test scores are obtained by summing the number of correctly answered test items.

5. Each examinee was assigned to a mastery state based on a comparison of his/her estimated domain score and the advancement score. Two assignments were made, one for each test administered.

6. The five summary statistics were calculated.

7. Steps 1 to 6 were repeated for each of five domain score distributions, and five test lengths (5, 10, 15, 20, and 40 test items). In addition, for two test lengths (5 and 10), the summary statistics were calculated for three advancement scores, one at 80%, and one below and another one above 80%.

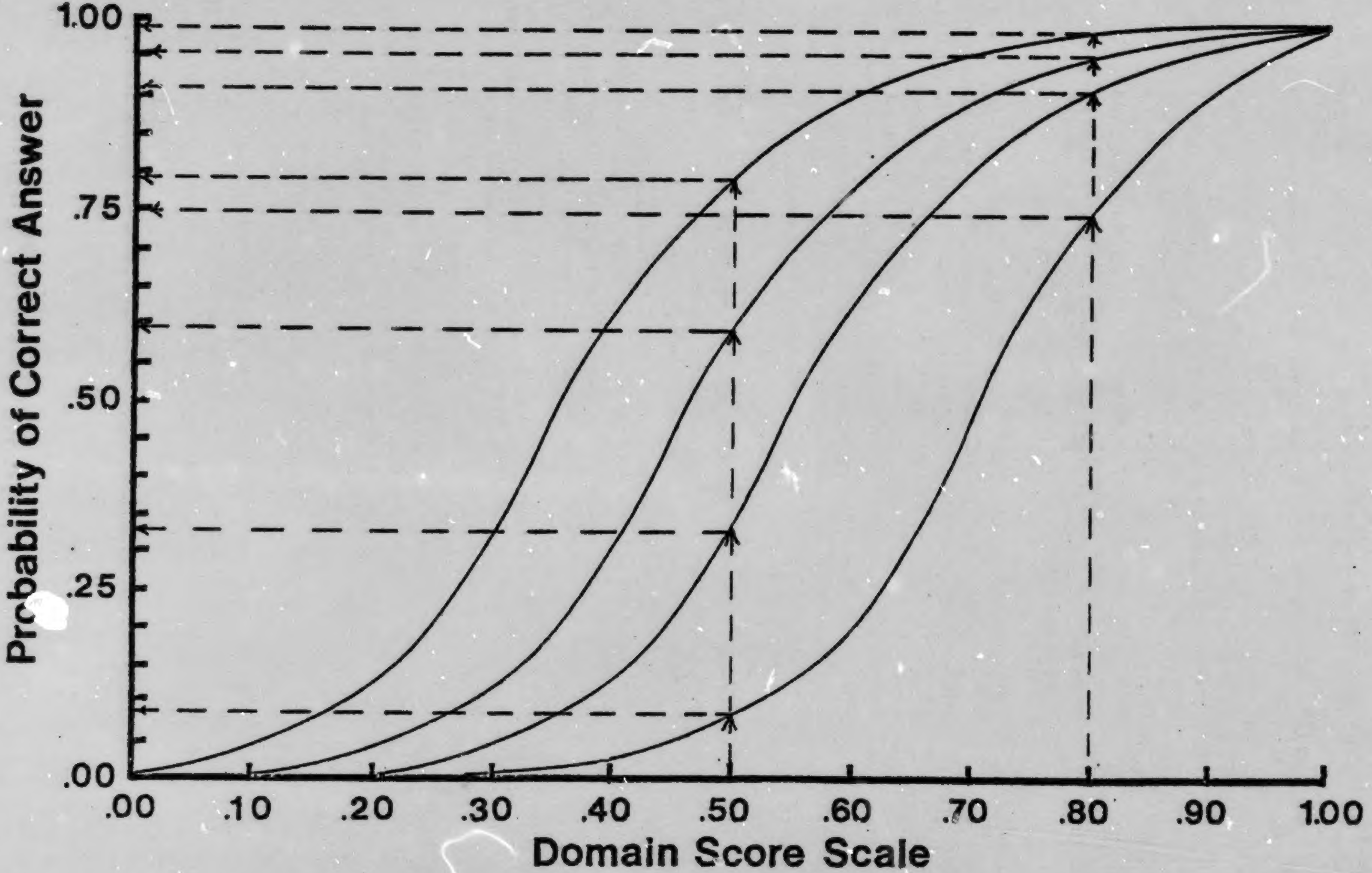


Figure 1. Item Characteristic Curves of Four Test Items and Probabilities of Correct Answers for Two Examinees

C. Results and Discussion

Effects of Test Length on Selected Test Score Reliability and Validity Indices

Figures 2 to 6 provide the relationships between test length and decision consistency, kappa, decision accuracy, predictive validity, and efficiency, respectively, for each of the five domain score distributions under consideration. In preparing the figures, statistical data were available for each of the domain score distributions at six test lengths: 0, 5, 10, 15, 20, and 40 items. Curves were drawn to be monotonically increasing, non-intersecting, and as close fitting to the data points as possible.

A number of observations and/or cautions concerning the use of Figures 2 to 6 are offered next:

1. Test score validity indices are lowest with homogeneous domain score distributions centered at or near a cut-off score. Domain score distribution one (and to a lesser extent) distribution two reflect this. The validity indices are highest for homogeneous domain score distributions where the center of the distribution is far from a cut-off score. These findings have several implications:

   (a) Short tests can be used when there is reason to believe that a group of examinees will do either very well or very poorly on a particular test. (Of course, if the prior belief about the distribution of domain scores is highly inaccurate, test score reliability and validity indices will be considerably lower than those predicted from the figures.)

2. Figures 2 to 6 apply to the case $\pi_o = \hat{\pi}_o = .80$. Such a situation is common in practice but variations in cut-off scores and advancement scores from .80 will reduce the usefulness of the results reported in the figures.

3. Details for using the figures in test development work will be offered later in the paper. It suffices to say here that the more important figures are those connecting test length to the validity indices. After an initial determination of test length has been made, Figures 2 and 3 can be used to predict test score reliability. If it is

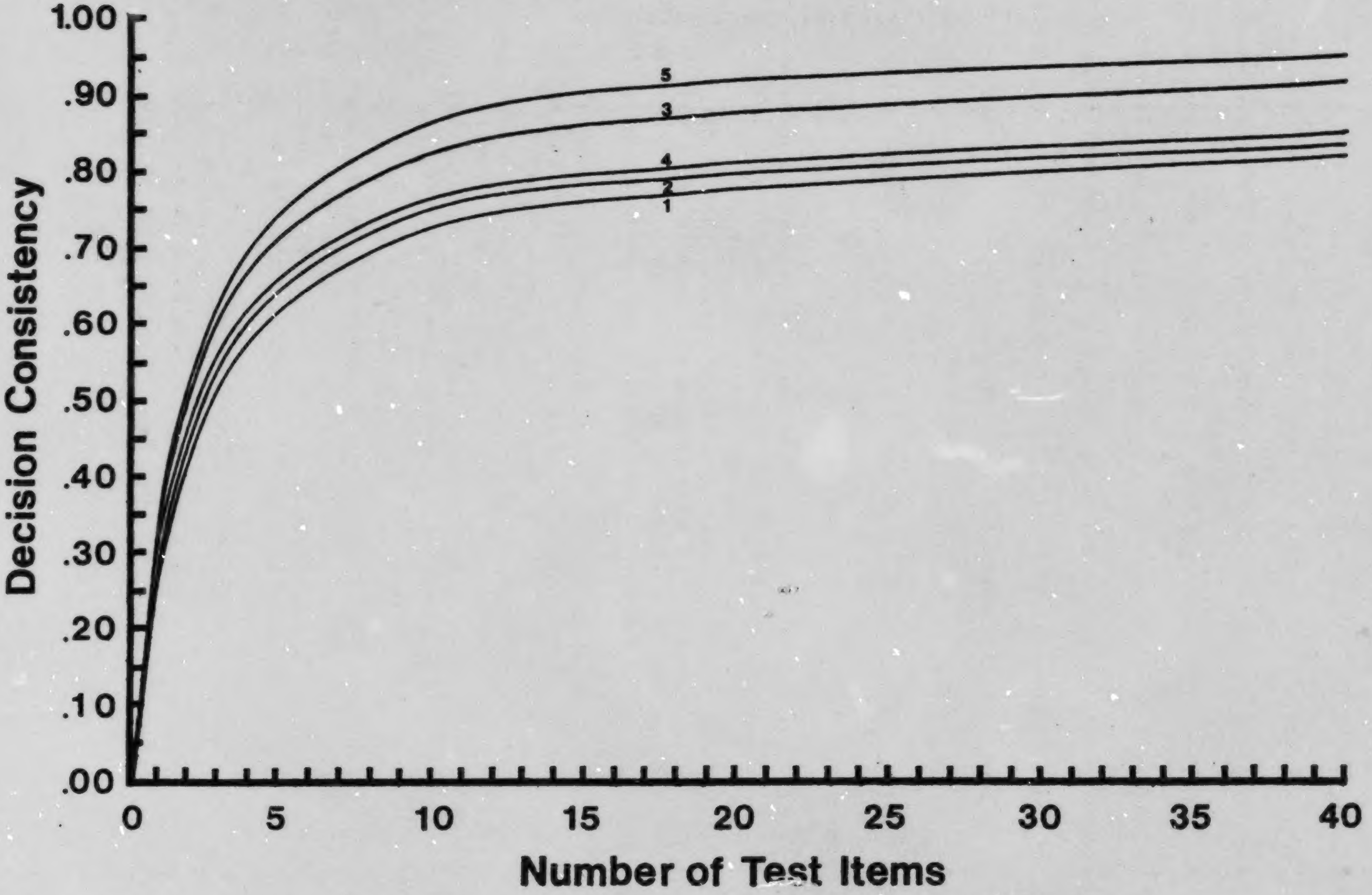


Figure 2. Relationship Between Decision Consistency and Test Length with Five Test Score Distributions

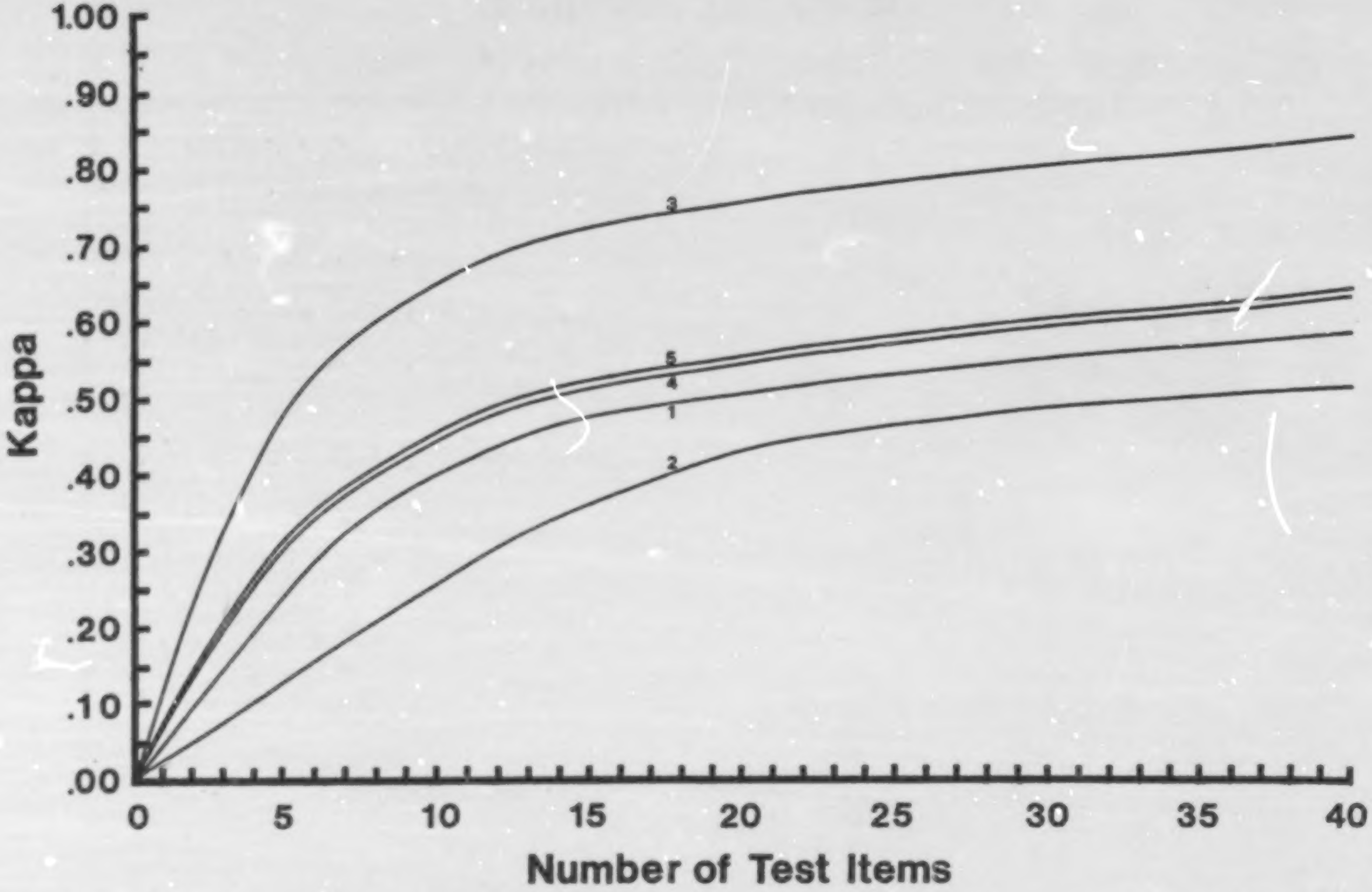


Figure 3. Relationship Between Kappa and Test Length with Five Test Score Distributions

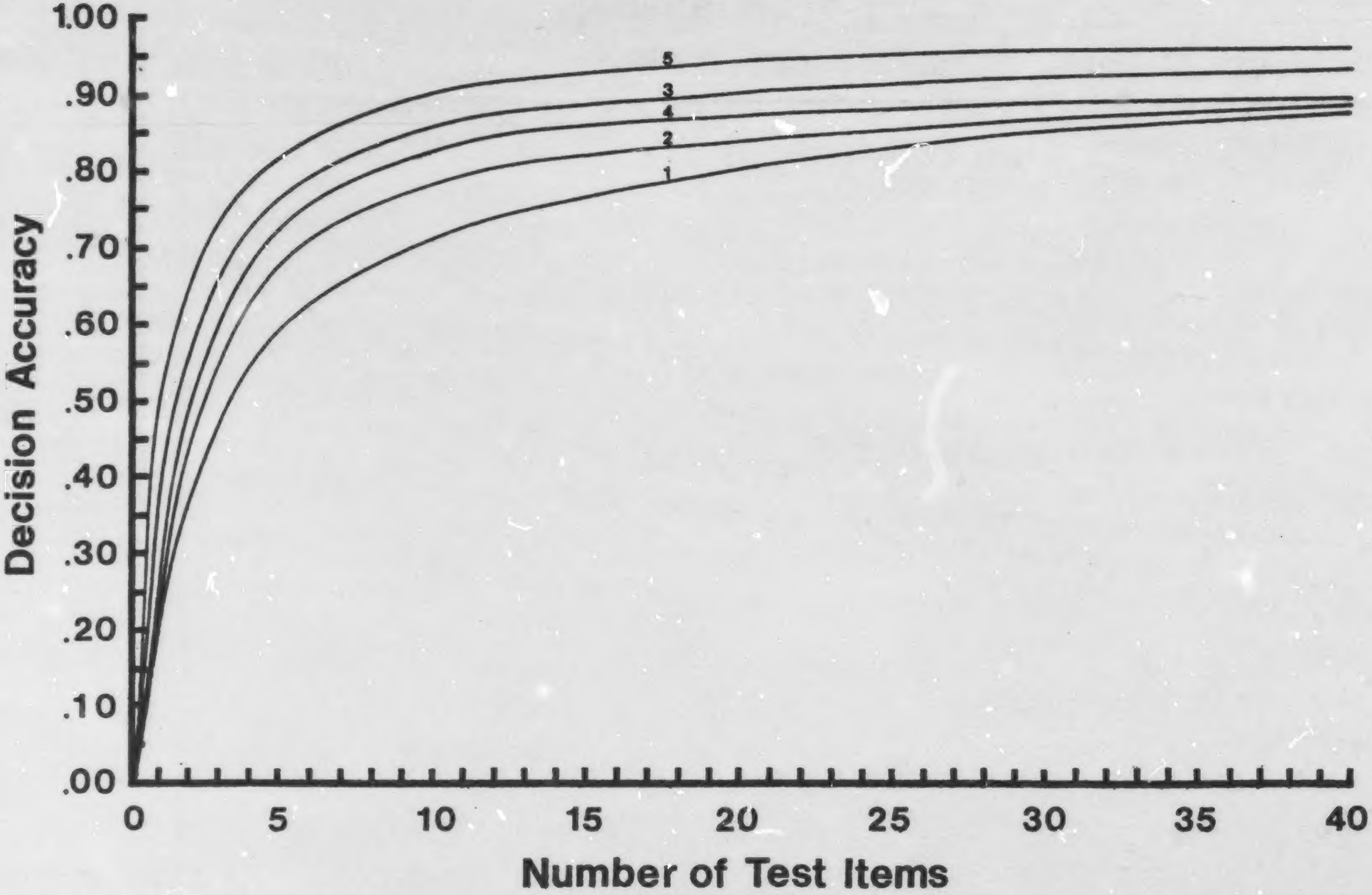


Figure 4. Relationship Between Decision Accuracy and Test Length with Five Test Score Distributions

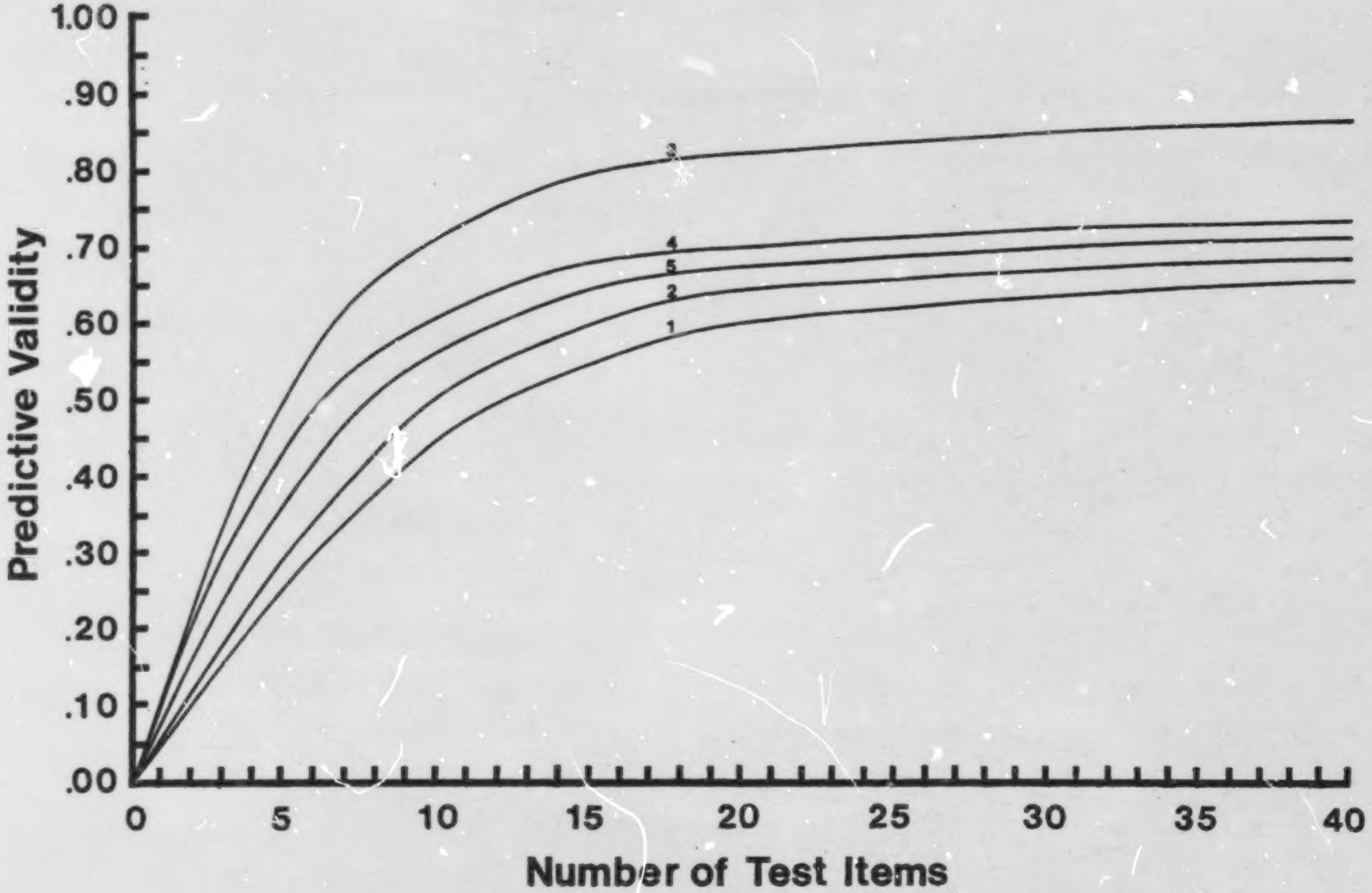


Figure 5. Relationship Between Predictive Validity and Test Length with Five Test Score Distributions

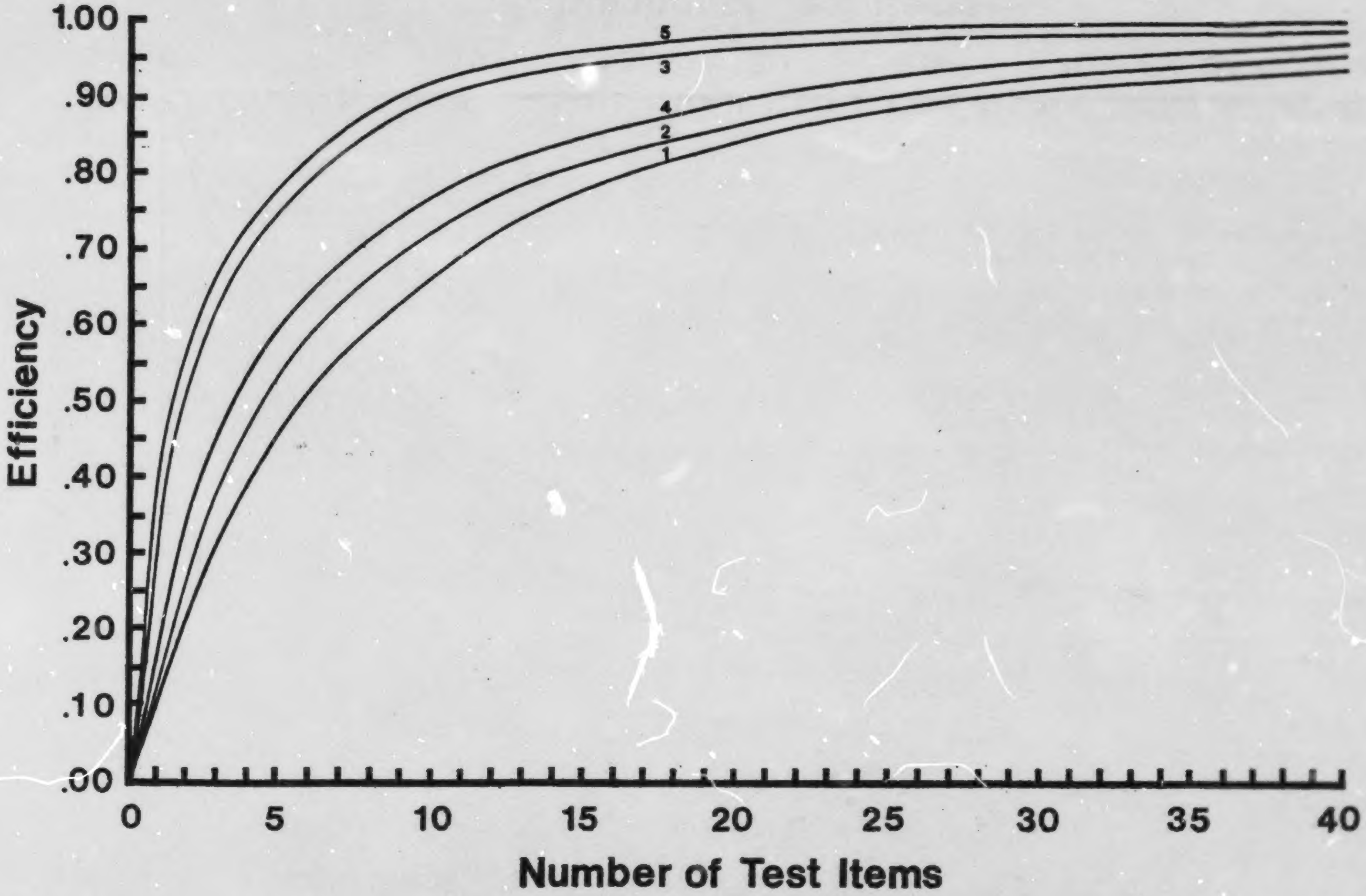


Figure 6. Relationship Between Efficiency and Test Length with Five Test Score Distributions

not high enough to meet some specified standard, the test plan must be revised to lengthen the required test.

4. Unfortunately, for many of the test lengths under consideration π_0 cannot equal $\hat{\pi}_0$ (for example with $\pi_0 = 80$, and an eight item test, $\hat{\pi}_0$ can be set equal to .75 or .875 but not .80). For test length and reliability results, the direction of errors in the figures will depend on the relation between $\hat{\pi}_0$ and the mean of the domain score distribution. Decision consistency is monotonically related to the difference between $\hat{\pi}_0$ and $\bar{\pi}$. The bigger the difference, the higher the value of decision consistency will be. On the other hand, for kappa, the highest values are obtained when $\hat{\pi}_0$ and $\bar{\pi}$ are fairly close.

   For test length and validity results, the direction of errors appears to depend in a complicated way on the relations among π_0, $\hat{\pi}_0$, and $\bar{\pi}$. More will be said about this in a later section.

Effects of Advancement Score on Test Score Reliability and Validity Indices

It was mentioned earlier that it is not always possible to set a cut-off score and an advancement score equal to the same value. Sometimes it is not even desirable to do so when the opportunity is available. For example, if false-positive errors are considerably more serious than false-negative errors, a test user may choose to set a very high advancement score and thereby minimize the number of false-positive errors. Such an action however will influence test score reliability and validity in a complex way. In this section of the paper a modest attempt is made to sort through a few of the complexities. Data on the reliability and validity indices for two test lengths, three advancement scores, and five domain score distributions are reported in Table 2. A few comments may help to interpret the results in the Table. Note, however, that because of sampling errors, not all of the results are consistent with the interpretations offered below.

Table 2

Effect of Advancement Score on Several Reliability and Validity Indices with Five Domain Score Distributions

Statistic	Test Length	Advancement Score	Domain Score Distribution				
			1	2	3	4	5
Decision Consistency	5	3	.72	.93	.84	.76	.71
	5	4	.64	.71	.76	.66	.71
	5	5	.74	.55	.76	.70	.87
	10	7	.73	.84	.80	.77	.73
	10	8	.74	.74	.81	.72	.86
	10	9	.77	.62	.84	.74	.89
Kappa	5	3	.22	.08	.58	.32	.41
	5	4	.28	.11	.49	.31	.35
	5	5	.24	.10	.49	.29	.34
	10	7	.47	.16	.51	.45	.30
	10	8	.46	.28	.62	.43	.47
	10	9	.33	.23	.67	.40	.40
Decision Accuracy	5	3	.43	.70	.72	.55	.62
	5	4	.60	.74	.82	.68	.76
	5	5	.83	.59	.82	.74	.88
	10	7	.56	.80	.75	.74	.78
	10	8	.69	.77	.87	.77	.90
	10	9	.83	.71	.89	.83	.95
Predictive Validity	5	3	.25	.09	.54	.36	.25
	5	4	.29	.32	.65	.40	.43
	5	5	.48	.22	.64	.48	.40
	10	7	.31	.38	.56	.54	.33
	10	8	.42	.51	.75	.55	.55
	10	9	.50	.42	.78	.58	.39
Efficiency	5	3	.02	.51	.55	.15	.51
	5	4	.45	.64	.78	.53	.75
	5	5	.82	.39	.83	.75	.93
	10	7	.43	.71	.81	.62	.79
	10	8	.66	.70	.89	.74	.93
	10	9	.83	.61	.93	.88	.97

Decision Consistency

It is very clear that as the advancement score moves away from the center of a domain score distribution, decision consistency increases. This explains why for the 10-item test and distribution five, decision consistency is lowest (.73) at $\hat{\pi}_0 = .70$ and highest (.89) at $\hat{\pi}_0 = .90$. The mean of the distribution is in the region of .60. The reverse result is obtained with distribution two. The highest value (.84) is obtained at $\hat{\pi}_0 = .70$ and the lowest value (.62) is obtained at $\hat{\pi}_0 = .90$. The mean of distribution two is about .90. Since the mean of distribution four is close to .80, it is not surprising to observe the lowest value (.72) at $\hat{\pi}_0 = .80$ and higher values at $\hat{\pi}_0 = .70$ (.77) and at $\hat{\pi}_0 = .90$ (.74).

Kappa

While the results are not too clear cut, it does appear that the highest values of kappa are obtained when an advancement score is near the middle of a domain score distribution. Huynh (1976) noted a similar finding in some of his work.

Decision Accuracy

Somewhat surprisingly, the value of decision accuracy is monotonically related to the distance between $\hat{\pi}_0$ and $\bar{\pi}$. The role that π_0 plays in the tabulated results is not readily apparent from the reported results.

Predictive Validity

There do not appear to be any trends in the results.

Efficiency

The results here are identical to those reported for decision accuracy and the explanation is the same.

Using the Results to Determine Test Length

Many factors will have an influence on the test length which is finally selected:

1. The shape (essentially variability) of the domain score distribution (regardless of which statistic is chosen, it is clear from Figures 2 to 6 that the variability of the

BEST COPY AVAILABLE

domain score distribution has a considerable influence on the results). In general, higher indices are obtained with heterogeneous domain score distributions.

2. The placement of cut-off scores (in general, higher validity indices are obtained if π_0 and π are not too close).

3. The selection of advancement scores (has a complicated relationship to test length).

4. The desired level of one of the reliability and/or validity indices (the higher the desired value, the longer the required test must be).

Six steps are offered next for determining test length in particular testing situations:

1. Select a primary statistic of interest (this is usually "decision accuracy").

2. Set a cut-off score (if $\pi_0 = .80$, proceed through the remaining steps; if $\pi_0 \neq .80$, it will be necessary to generate additional results using the method described in the last section of this paper).

3. Set advancement scores corresponding to test lengths under consideration which are near .80 (if $\hat{\pi}_0 \doteq .80$ Figures 2 to 6 will provide usable results).

4. Specify a prior belief about the domain score distribution for the group of examinees who will be assessed. If conservative results are desired, it is best to work with homogeneous distributions centered around $\pi_0 = .80$.

5. Choose (a) or (b)

   (a) With the statistic identified in step 1, and a desired value for the statistic, find the correct figure and read off the corresponding test length from the curve corresponding to the domain score distribution selected in step 4.

   For example, suppose a test developer desired a decision accuracy statistic equal to .80 and the most likely domain score distribution is number 1. From Figure 4, the corresponding test length is 21 items.

   (b) With the reliability or validity statistic selected in step 1, and several test lengths of interest, find the corresponding values of the desired statistic for the test lengths of interest. Select the test length which seems suitable.

6. Check "decision consistency" and/or "kappa" for the test length selected in step 5. (With the example in 5a above, the value is .75 for decision consistency.) If the value is too low for the intended purpose of the test, determine a value which is not, read off the corresponding test length, and then repeat step 5a or 5b again.

The values provided in the figures are only approximations. Still, they should be helpful to test developers who aspire to set their test lengths in a way which is not totally dependent on guess work.

D. Suggestions for Further Research and Development

Because of (1) the considerable importance of the topics under study in this paper, and (2) the paucity of practical research results, it is easy to suggest many directions for further work. For one, a computer program is needed into which a test developer can (a) provide a prior belief about the shape of a domain specification distribution for some group of examinees to be tested, (b) select a test model (probably the binomial or the compound binomial), (c) select one or more reliability and validity indices of interest, and (d) select test lengths and advancement scores of interest. The output from the computer program would provide a basis for determining test length.

One of the spin offs from this simulation study is the availability of a computer program that has some of the features mentioned above. It can be used by test practitioners to generate additional results to those reported in the paper. Practitioners must only specify (1) a prior belief about the distribution of domain scores, (2) suggest test lengths, cut-off scores, and advancement scores, and

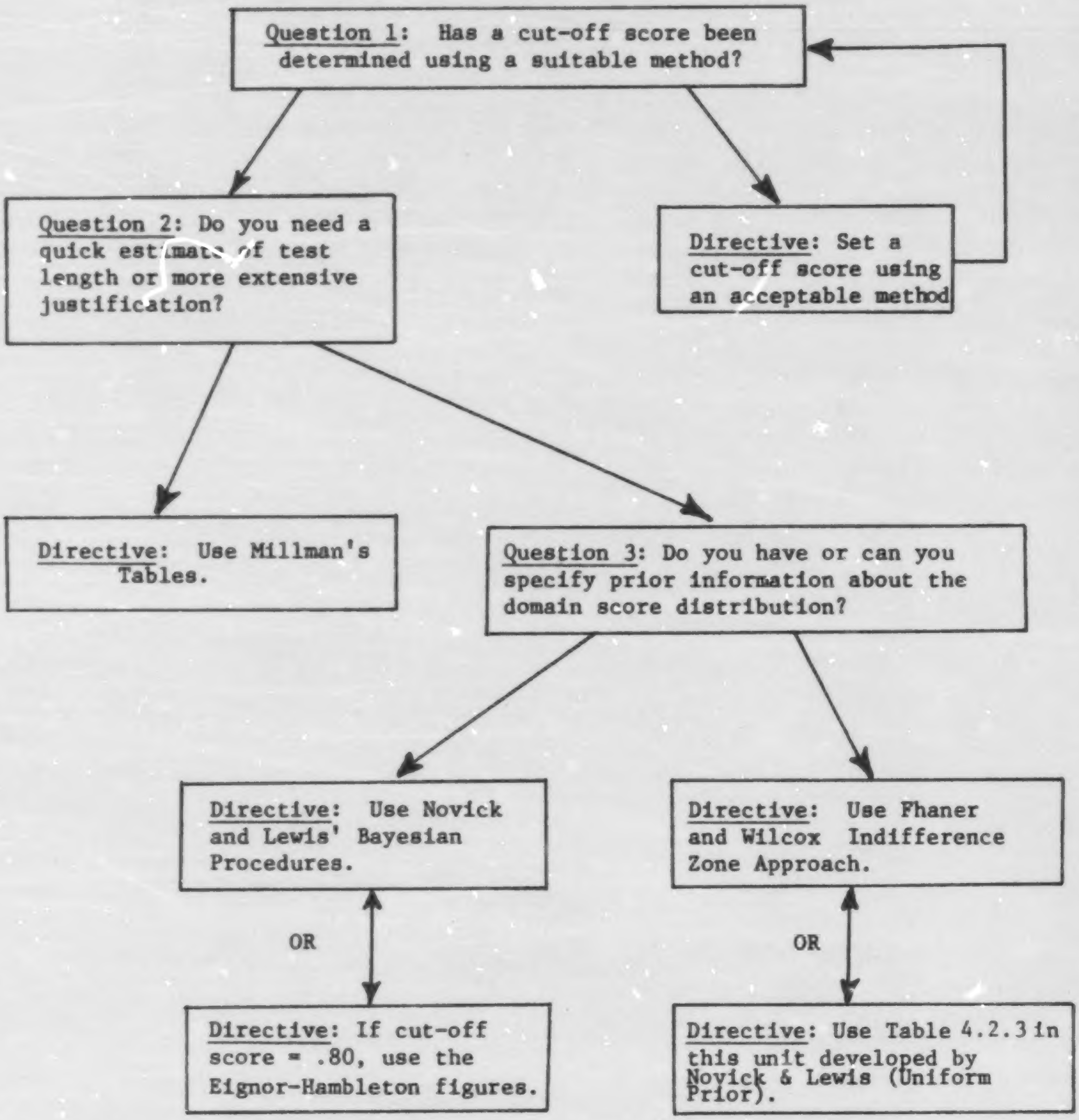
(3) select either the binomial or compound binomial test model from which to simulate examinee item response data. Figures similar to those reported in this study can be quickly obtained. A write-up of the current computer program is in preparation and will be available soon. One drawback is that it is not as easy a system to use nor does it have as many features as might be desirable.

A second area for further work is in the area of "guidelines for interpreting the reliability and validity indices." In the area of norm-referenced testing, even with a plethora of textbooks available and the training many people have had, there is still considerable confusion about the correct interpretations of reliability and validity indices. Because of the newness of the five statistics used in this study, it seems clear that if they are to have any value at all, increased effort must be given to training test developers in the use of these and other relevant statistics.

Third, the validity of the relationships reported in Figures 2 to 6 among test length, cut-off scores, advancement scores, and domain score distributions, and five reliability and validity indices should be compared to existing results reported on real test data. In a very limited way, some of the theoretical results reported in this paper were compared to results obtained from real data. The differences were very small but considerably more work of this general type should be done. The reliability results would be particularly easy to check. Only the examinee responses to large sets of test items keyed to objectives would be required. "Tests" of varying lengths could be drawn from the examinee-item pool of data keyed to a particular objective, "parallel-forms" constructed, and various advancement scores considered. Via the method of sampling of examinees, assuming the "pool" of examinees was heterogeneous and large enough, nearly any domain score distribution could also be studied as well.

4.2.7 Method of Selecting a Procedure For Determining Test Length

Answers to the three questions below will provide a basis for selecting one of the four methods of determining test length.



Question 1: Has a cut-off score been determined using a suitable method?

Directive: Set a cut-off score using an acceptable method

Question 2: Do you need a quick estimate of test length or more extensive justification?

Directive: Use Millman's Tables.

Question 3: Do you have or can you specify prior information about the domain score distribution?

Directive: Use Novick and Lewis' Bayesian Procedures.

Directive: Use Fhaner and Wilcox Indifference Zone Approach.

Directive: If cut-off score = .80, use the Eignor-Hambleton figures.

Directive: Use Table 4.2.3 in this unit developed by Novick & Lewis (Uniform Prior).

4.3 Test Item Selection

The item selection process is quite simple provided the criterion-referenced test constructor has been careful in defining the domain of concern and in constructing test items (see Unit 2). That is, the test developer has to have been careful to define the size of his/her domain to be consonant with the test's purpose. If the purpose of testing is to make major level decisions on, for instance the school level, a large domain size can be tolerated. If, however, the purpose of testing is to provide information for remedial instruction, a smaller domain size is needed. Popham (1978) has offered some suggestions for ascertaining domain size. The critical point for item selection is that the domain be a reasonable size so that proper sampling from the domain can occur. If the domain is so large that it is difficult to see how to generate a set of items from the domain for the test, then the domain must be broken up into sub-domains and items generated for those sub-domains. The sampling process should be clear for these sub-domains. Thus, it is critical that the domain be of a size that a set of items can be clearly constructed from the domain, and then the sampling process can be carried out without complications.

Having defined a domain size that is manageable for sampling is not enough; the test developer must also be careful to ascertain that all the items constructed for the domain do indeed "tap" the behavior specified. The items must adhere to the restrictions imposed on the domain specifications.

If the size of the domain is manageable for the sampling process and the test developer is sure that the items generated "tap" the specified behavior, then the item selection process is quite simple.

The test is constructed by taking either a random or stratified random sample of items from the domain. It should be noted that if the domain has been explicitly defined (see section 2.1 of Unit 2), then a random sample of items can be taken. If the domain has to be defined implicitly, as is the case with domain specifications, then only a representative set of items defining the domain has been generated, and a random sample is drawn from that set for the test. That is really a technical distinction referring to the domain; in either case, the items should (in theory) be selected randomly for the test.

A word of caution should be presented at this point. Unlike the procedures for norm-referenced tests, statistical indices should not be used in the item selection process. Item difficulty and item discrimination arenot useful in the item selection process; these indices may be useful in helping to detect flawed items in the item validation stage (see Unit 3). According to Millman (1974):

> Selection of items on these criteria can result in a test where the items are not representative of the domain in difficulty level or in the underlying attributes being measured. An examinee's status relative to a well-defined domain can best be gleaned from the examinee's responses to a representative sample of items from the item population. Items chosen by empirical means are likely to be average in difficulty and more homogeneous than is true for all the items. The use of item statistics destroys the random selection process, a defining characteristic of [criterion-referenced tests]. Unless items are selected randomly, the estimate of a person's domain score loses meaning and the interpretability of the test score is reduced.

In sum, items should be selected by random sampling from the complete set of items generated for domains defined explicitly or from the representative set of items generated for domains defined implicitly.

One advantage of choosing representative sets of test items is that examinee test scores (or proportion-correct scores) provide "unbiased" estimates of their domain scores. It is possible also to set standards and interpret examinee test performance relative to those standards. Unfortunately, when the number of test items is small (as is frequently the case), the consistency of decisions (competent/incompetent) across a retest administration or across a parallel-form administration of a test may be distressingly low. Increasing the number of test items measuring each objective is helpful but often it is not feasible to do so. One answer to the dilemma is as follows: When the primary purpose of the testing program is to make dichotomous decisions about examinees, a more effective test can be produced if test items from the available pool of test items measuring each objective are selected based on their statistical properties. Specifically, if (say) a standard is set at 80%, it would be best to select test items which have p-values (item difficulty levels) in the region of .80 and which have the highest discrimination indices. A test constructed in this way will have maximum discriminating power in the region where decisions are being made and therefore more reliable and valid decisions will result. One possible drawback is that scores derived from the test cannot be used to make descriptive statements about examinee levels of performance on the objectives measured in the test. This is because test items measuring each objective will not usually constitute a representative sample. In theory, there is at least one way to make descriptive statements about examinee levels of performance on the objectives measured by a test when non-random or non-representative samples of test items are chosen. It can be done by introducing concepts and models from the field of latent trait theory. The feasibility, however, of such an approach has not been tested.

4.3.1 Post Item Selection Checklist

In Unit 2 of these materials a set of checklists were offered that should be useful at the item writing stage of the criterion-referenced test development process. Most of the questions posed in those checklists can be answered after the items are written; there are, however, a number of questions that can only be answered after the test items have been selected and organized into a test. The checklist that follows presents the questions that are appropriate to ask after the test items have been selected and assembled in a test.

Test Directions and Item Selection

Review Form

Domain Specification: ______________ Reviewer: ___________ Date: __________

Test Directions	Yes	No	Unsure
1. Do the directions indicate the test's purpose?	✓		
2. Do the directions indicate how the test items will be scored?	✓		
3. Do the directions indicate how examinees are to "mark" their answers (on the test booklet or a separate answer sheet)?	✓		
4. Are there any practice test items?	✓		
5. Do the directions indicate the time allowed to complete the test items?	✓		
Test Items			
6. Do the test items represent at least an adequate sample from the domain of items defined by the domain specification?	✓		
7. Do any of the test items contain clues which may help examinees answer other test items measuring the domain specification?		✓	
8. Will examinees learn anything from one or more test items which will help them answer other test items?		✓	
9. Have the items been checked with content and measurement specialists to try and eliminate ambiguity, technical errors, and other errors in item writing?	✓		
10. Has the number of item formats been kept to a minimum?	✓		

	Yes	No	Unsure
11. Were the most "valid" item formats used?	✓		
12. Were items in the same format grouped together?	✓		
13. Do the correct answers follow essentially a random pattern?	✓		

Multiple-Choice

	Yes	No	Unsure
14. Has the number of negatively stated item stems been kept to a minimum (less than 10%)?	✓		

True-False

	Yes	No	Unsure
15. Are the true statements of the same length as the false statements?	✓		

"✓" indicates the desired response.

4.4 Preparation of Directions

In this and subsequent sections of this unit, the procedures to be described for criterion-referenced test development are essentially the same for norm-referenced test development. Because such procedures are well-documented, what follows are some helpful hints for the reader, along with the listing of references that may be referred to for a more in-depth discussion.

Payne (1974) has presented seven criteria that should be kept in mind when writing test directions. These criteria are from the Traxler (1951) paper. These are:

1. Assume that the examinees and examiner know nothing at all about objective tests.
2. In writing the directions, use a clear, succinct style. Be as explicit as possible, but avoid long drawn-out explanations.
3. Emphasize the more important directions and key activities through the use of underlying, italics, or different type size or style.
4. Give the examiner and each proctor full instructions on what is to be done before, during, and after the administration.
5. Field or pretest the directions with a sample of both examinees and examiners to identify possible misunderstanding and inconsistencies and gather suggestions for improvement.
6. Keep the directions for different forms, subsections, or booklets as uniform as possible.
7. Where necessary or helpful, give practice items before each regular section.

Gronlund (1976) states that while the directions should be as simple and concise as possible, they must contain information on each of the following:

1. The purpose of the test.
2. The time allowed to complete the test.
3. How answers should be recorded (on the test itself or a separate answer sheet).
4. Whether or not to guess when in doubt about the answer.

Gronlund (1976) has an excellent discussion of these four areas of concern. Ahmann and Glock (1975) also have a good discussion on preparation of directions.

In reference to Gronlund's fourth point, about guessing on criterion-referenced tests, some helpful comments can be made at this point, both about the guessing itself and whether or not to use correction for guessing formulas. First of all, it is unlikely that in a criterion-referenced testing context a student would be guessing blindly at an answer. When these tests are used in instructional settings, such as after a unit of study, the student is likely to have partial knowledge about an answer if he/she does not know the answer. The guideline is if the student can eliminate any of the response options on a test question, he/she should be encouraged to attempt the question utilizing the smaller option set. Hence, in such a case, the student should be encouraged to attempt the item.

Correction-for-guessing formulas have been utilized in norm-referenced testing situations because of the concern that the proper rank-ordering of students based on test results may be upset due to the predisposition of certain students to guess randomly at questions, and other examinees to omit questions even when they are reasonably certain of their answers. Further, it is known that if all examinees have sufficient time to answer all items, there is no difference in the rankings of students on corrected scores as compared with uncorrected scores. We feel that correction-for-guessing formulas are suitable only for the norm-referenced context, where rank-ordering is the concern. Further, it would make little sense to use them for criterion-referenced tests because with these tests students are usually given sufficient time to complete the questions.

4.5 Layout and Test Booklet Preparation

In assembling the items into a test, a decision must be made concerning the best item arrangement. There are two possible ways of organizing a set of items in a criterion-referenced test:

1. If there are multiple item types, the items should be arranged so that all items of the same type are grouped together.
2. For many purposes, it may be desirable to group together items that measure the same objective or domain generated from a domain specification.

In only certain situations will both possibilities be able to be applied simultaneously. Usually one method of organization will be chosen over the other, and this will depend upon the purpose for testing. For instance, if the test is being used to diagnose problems for subsequent assignment of students to remedial activities, the test developer would probably want to group items tapping the same objective together. This would give an immediate indication of those objectives the student is having difficulty with. Further, it may be possible to organize by item type within objective if there are a large number of test items per objective. In sum, the choice of method of organization will depend upon the purpose for testing.

Gronlund (1976) suggests that if the organization is by item type, because certain item types are more difficult than others and the simpler activities should come first, the following order should be used:

1. True-false items
2. Matching items
3. Short-answer items
4. Multiple-choice items
5. Essay questions

He further suggests that organization by item type should always be considered first, and that only in certain situations should alternate organizational schemes be considered. According to Gronlund (1976):

> This arrangement provides for the finest set of directions; it is easier for the pupils since they can retain the same mental set throughout each section; and it greatly facilitates scoring.

The following guidelines offered for test booklet preparation are relevant for teacher prepared tests. These points have been synthesized from Gronlund (1976) and Noll and Scannell (1972). An indepth discussion of procedures for preparation and reproduction of the test can be found in an article by Thorndike (1971). This is the most recent, indepth discussion of these procedures that the authors have seen.

In these materials, the following useful guidelines are offered:

1. Make sure that test items are spaced so that they can be read, answered, and scored with the least amount of difficulty. Double space between items.
2. Make sure all items have generous borders.
3. Multiple-choice items should have the alternatives listed vertically beneath the stem.
4. Do not split an item onto two separate pages.
5. With interpretation exercises, place the introduction on a facing page with all items referring to it on a single page.
6. If not using an answer sheet, the space for answering should be down the left side of the page.
7. The most convenient method of response is circling correct answers.
8. Test items should be numbered consecutively throughout the test.
9. Tests reproduced by processes available to school systems should be duplicated on one side of the sheet only.
10. If a separate answer sheet is used, test booklets can be reused. They should be numbered so a check can be made for a complete set of materials after test administration.

4.6 Preparation of Scoring Keys

If a standard, commercial answer sheet is used, either the answer sheets can be scored by machine or a punch-out overlay template can be used in scoring. If a hand-scoring answer key is to be used, Payne (1974), based on the Traxler (1951) article, describes three varieties of hand-scoring keys that can be useful. These are: The fan or accordian, strip, and cut-out keys. What follows is a brief description of each. The descriptions are taken directly from Payne (1974):

Fan Key: This key consists of a series of columns, extending from the top to the bottom of the page, on which are recorded acceptable answers or directions scored for the individual items. The key and the answer sheet are the same size and identically spaced. Usually each column corresponds to a page of the test. The key is folded along vertical lines separating its columns and is superimposed on the appropriate page of the test or next to the appropriate column of the answer sheet and matched to the corresponding responses.

Strip Key: Similar to the fan key, this method employs the use of separate columns, usually on cardboard.

Cut-Out Key: Windows are cut out to reveal letters, numbers, words, or phrases on the answer sheet. The key is superimposed on a page of the test or answer sheet.

Gronlund (1976) offers some helpful hints that can be used in the actual scoring process. A most useful hint is to draw a red line through the correct answers of items missed rather than through the wrong answers. This indicates to the student which items he/she missed and at the same time indicates the correct answer.

4.7 Preparation of Answer Sheets

If a teacher prepared answer sheet is to be used, the following simple guidelines may be helpful:

1. Make sure that the number on the items correspond with the numbers on the answer sheet.
2. Number the items on the answer sheet consecutively down the pages rather than across.
3. Make all lines for answers exactly the same length.

If a commercially prepared answer sheet is to be used, the following suggestions may be helpful:

1. Make sure that the answer sheet does not have more response options than the test.
2. Try to obtain answer sheets that have answer spaces running down a column of the answer sheet rather than across. If the answer spaces run across, make sure to notify the students.
3. Try to purchase an answer sheet that has approximately the same number of answer spaces as questions on the test.

4.8 Test Administration

An excellent discussion of factors of concern in the test administration process is contained in an article by Clemans (1971). What follows is material discussed in the Payne (1974) book and in Gronlund (1976).

In order to insure optimal conditions, so that test scores can have meaning, Prescott has prepared the following set of guidelines for administration before, during, and after the test (taken from Payne, 1974). These guidelines are relevant for both standardized and classroom tests.

Before the Testing Date

1. Understand nature and purposes of the testing:
    a. Tests to be given.
    b. Reasons for giving tests.
2. Decide on number to be tested at one time.
3. Decide on seating arrangements.
4. Decide on exact time of testing.
    a. Avoid day before holiday.
    b. Avoid conflicts with recess of other groups.
    c. Make sure there is ample time.
5. Procure and check test materials:
    a. Directions for administering.
    b. Directions for scoring.
    c. Test booklets:
        (1) One for each pupil and examiner.
    d. Answer sheets:
        (1) One for each pupil and examiner.
    e. Pencils (regular or special).
    f. Stopwatch or other suitable timer.
    g. Scoring keys.
    h. "Testing—Do Not Disturb" sign.
    i. Other supplies (scratch paper, etc.).
6. Study test and directions carefully.
    a. Familiarize yourself with:
        (1) General make-up of test.
        (2) Time limits.
        (3) Directions.
        (4) Method of indicating answers.
    b. Take the test yourself.

7. Arrange materials for distribution.
   a. Count number needed.
8. Decide on order in which materials are to be distributed and collected.
8. Decide what pupils who finish early are to do.

Just Before Testing

1. Make sure central loudspeaker is disconnected.
2. Put up "Testing—Do Not Disturb" sign.
3. See that desks are cleared.
4. See that pupils have sharpened pencils.
5. Attend to toilet needs of pupils.
6. Check lighting.
7. Check ventilation.
8. Make seating arrangements.

During Testing

1. Distribute materials according to predetermined order.
2. Caution pupils not to begin until you tell them to do so.
3. Make sure that all identifying information is written on booklet or answer sheet.
4. Read directions exactly as given.
5. Give signal to start.
6. Write starting and finishing times on the chalkboard.
7. Move quietly about the room to:
   a. Make sure pupils are marking answers in the correct place.
   b. Make sure pupils are continuing to the next page after finishing the previous page.
   c. Make sure pupils stop at the end of the test.
   d. Replace broken pencils.
   e. Encourage pupils to keep working until time is called.
   f. Make sure there is no copying.
   g. Attend to pupils finishing early.
8. Permit no outside interruptions.
9. Stop at the proper time.

Just After Testing

1. Collect materials according to predetermined order.
2. Count booklets and answer sheets.
3. Make a record of any incidents observed that may tend to invalidate scores made by pupils.

In addition to these guidelines, Gronlund (1976) offers the following four suggestions about activities to avoid when administering the test:

1. Do not talk unnecessarily before the test.
2. Keep interruptions during the test to a minimum.
3. Avoid giving hints to pupils who ask about individual items.
4. Prevent cheating, if necessary.

Further, to present undue test anxiety, Gronlund (1976) suggests that the teachers or test administrator be careful not to:

1. Threaten pupils with a test if they are not behaving.
2. Warn pupils to do their best "because the test is important."
3. Tell students they must work fast to complete the items on time.
4. Threaten unpleasant activities if they fail.

In sum, these guidelines for administering a test should aid in assuming that all the students being tested are being given a fair chance to demonstrate what they know on the domains being tested.

4.9 References Cited

Ahmann, J. S., & Glock, M. D. *Evaluating pupil growth*. (5th ed.) Boston: Allyn and Bacon, 1975.

Clemans, W. V. Test administration. In R. L. Thorndike (Ed.). *Educational Measurement*. (2nd ed.) Washington: American Council on Education, 1971.

Eignor, D. R., & Hambleton, R. K. Effects of test length and advancement score on several criterion-referenced test reliability and validity indices. *Laboratory of Psychometric and Evaluative Research Report No. 86*. Amherst, MA: School of Education, University of Massachusetts, 1979.

Fhaner, S. Item sampling and decision-making in achievement testing. *British Journal of Mathematical and Statistical Psychology*, 1974, *27*, 172-175.

Gronlund, N. E. *Measurement and evaluation in teaching*. (3rd ed.) New York: MacMillan, 1976.

Gronlund, N. E. *Constructing achievement tests*. (2nd ed.) Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Hambleton, R. K., & Eignor, D. R. Adaptive testing applied to hierarchically-structured objectives-based curricula. In D. J. Weiss (Ed.), *Proceedings of the 1977 Computerized Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota, 1978.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test Scores*. Reading, Mass.: Addison-Wesley, 1968.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, California: McCutchan Publishing Co., 1974.

Noll, V. H., & Scannell, D. P. *Introduction to educational measurement*. (3rd ed.) Boston: Houghton Mifflin, 1972.

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Payne, D. A. The assessment of learning: Cognitive and affective. Lexington, MA: D. C. Heath, 1974.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Prescott, G. A. Test service bulletin 102, Test administration guide. New York: Harcourt Brace Jovanovich, Undated.

Spineti, J., & Hambleton, R. K. A computer simulation study of tailored testing strategies for objectives-based instructional programs. Educational and Psychological Measurement, 1977, 37, 139-158.

Thorndike, R. L. Reproducing the test. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Traxler, A. E. Administering and scoring the objective test. In E. F. Lindquist (Ed.), Educational measurement. Washington: American Council on Education, 1951.

Wilcox, R. A note on the length and passing score of a mastery test. Journal of Educational Statistics, 1976, 1, 359-364.

References Cited in the Eignor-Hambleton Paper

Algina, J., & Noe, M. J. A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 1978, 15, 101-110.

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9.

Block, J. H. Student learning and the setting of mastery performance standards. Educational Horizons, 1972, 50, 163-190.

Eignor, D. R. Psychometric and methodological contributions to criterion-referenced testing technology. Unpublished doctoral dissertation, University of Massachusetts, 1979.

Fhaner, S. Item sampling and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Hambleton, R. K., & Eignor, D. R. A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70. Amherst, MA: School of Education, University of Massachusetts, 1978.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R. K., & Rovinelli, R. A Fortran IV program for generating examinee response data from logistic test models. Behavioral Science, 1973, 17, 73-74.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Livingston, S. A. Assessing the reliability of tests used to make pass/fail decisions. COPA Research Report. Princeton, NJ: Educational Testing Service, 1978.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

Marshall, J. L., & Haertel, E. H. The mean split-half coefficient of agreement: A single administration index of reliability for mastery tests. Unpublished manuscript, University of Wisconsin, 1976.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, CA: McCutchan Publishing Co., 1974.

Novick, M. R., & Jackson, P. H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.), *Problems in criterion-referenced measurement*. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, Univesrity of California, 1974.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1-9.

Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 1976, *13*, 265-275.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 1978, *15*, 111-116. (a)

Subkoviak, M. J. The reliability of mastery classification decisions. Paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, October 27, 1978. (b)

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*, 263-268.

Wilcox, R. A note on the length and passing score of a mastery test. *Journal of Educational Statistics*, 1976, *1*, 359-364.

Wilcox, R. R. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: An empirical Bayes approach. *Journal of Educational Statistics*, 1977, *2*, 289-307.

Unit 5

Reliability, Validity and Norms

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

Table of Contents

Page

5.0 Overview of the Unit. 1

5.1 Criterion-Referenced Test Score Uses. 2

5.2 Approaches to Reliability Assessment. 3

5.2.1 Early Work. 3
5.2.2 Reliability of Domain Score Estimates 7
5.2.3 Reliability of Mastery Classification Decisions . . . 13
5.2.4 Summary of the Reliability Discussion 26

5.3 Validity of Criterion-Referenced Tests. 28

5.3.1 Introduction. 28
5.3.2 Clarification of Several Validity Issues. 29
5.3.3 Content Validation Studies. 32
5.3.4 Construct Validation Studies. 35

Guttman Scalogram Analysis. 36
Factor Analysis 37
Experimental Studies of Sources of Invalidity 38

5.3.5 Summary 39

5.4 Norms for Interpreting Criterion-Referenced Test Scores . . 40

5.5 References. 44

5.0 Overview of the Unit[1]

This unit covers step ten of the Criterion-Referenced Test Development and Validation Model presented in Unit 1.

A good test, whether it is norm-referenced or criterion-referenced, must result in reliable and valid test scores. The particular form these two psychometric concepts take (i.e., how they will be estimated) will depend on the intended use of the criterion-referenced test scores. In this unit of the materials, we will offer procedures for ascertaining the reliability and validity of criterion-referenced test scores.

If the procedures discussed in Units 2, 3, and 4 are carefully followed, a criterion-referenced test score can give detailed information on what an individual can and can't do with respect to a content domain. Sometimes this information isn't enough however; for instance, a decision-maker might also want to know how well a student (or group of students) is performing relative to the performance of other groups (perhaps last year's graduating class or a group of students in a neighboring school district). Norms data can supply the extra information necessary for the decision maker to determine how well, on a comparative basis, an individual (or group) is performing. In section 5.4, we will discuss the use of norms with criterion-referenced tests.

---

[1] Several sections of the material are from Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

5.1 Criterion-Referenced Test Score Uses

Two uses of criterion-referenced test scores are of special interest in our work. The first involves the estimation of examinee domain scores. In this application, it is important to minimize the "error" defined as the difference between as examinee's "estimated domain score" and "domain score." The second application involves the assignment of examinees to mastery states (where each state is "keyed" to an instructional decision) based on their criterion-referenced test score performance. In this second application, among other things, the test score user must be concerned about the "errors" arising from inconsistent mastery state assignments across parallel-form administrations of the test or across a retest administration of the test.

5.2 Approaches to Reliability Assessment

5.2.1 Early Work

Perhaps the first discussion of the reliability of criterion-referenced tests was by Popham and Husek (1969). These authors took the point of view that while internal consistency and temporal stability may be important characteristics of test scores that result from criterion-referenced measurement, the coefficients prescribed by classical test theory for assessing these characteristics may be inappropriate. They noted the well-known result that test score reliability for a group of examinees is dependent on test score variability. Since it is not uncommon to observe rather homogeneous distributions of criterion-referenced test scores, they feared that test developers would "scrap" their tests because of low reliability values. Basically, they argued that test developers should not worry too much if they obtained low classical reliability estimates (low values were to be expected). They did not, however, suggest any concrete alternate approaches for estimating reliability of criterion-referenced tests. In hindsight, they might have suggested that test developers "create" test score variance by "pooling" test performance of two groups of examinees--those expected to be "masters" of the material included in a test (perhaps a group of examinees after instruction) and those who would be expected to be "non-masters" (perhaps a group of examinees prior to receiving instruction). It then would be possible to apply any of the classical reliability approaches and interpret the results in the usual way (see, for example, Haladyna, 1974). On the other hand, there would still remain problems in using classical approaches to reliability with criterion-referenced tests. These problems will be discussed below.

Hambleton and Novick (1973) also addressed the matter of classical test theory applications to criterion-referenced tests. They noted:

> Thus, it seems clear that the classical approaches to reliability and validity estimation will need to be interpreted more cautiously (or discarded) in the analysis of criterion-referenced tests. Perhaps, an even more serious reservation concerning the classical approach to reliability and validity estimation for criterion-referenced tests, if one looks at these psychometric concepts in decision-theoretic terms, is that the correlational method represents an inappropriate choice of a loss function (squared-error loss in the π metric) with which to evaluate a test (p. 167).

The latter of their two points is important but unfortunately not often cited by test developers as a reason for seeking out new testing methods for the design, interpretation, and use of criterion-referenced tests.

One of the first suggestions for an approach to the reliability of criterion-referenced tests came from Livingston (1972a). He began his interesting work by assuming that the purpose of a criterion-referenced test was to discriminate each examinee's estimated domain score from a cut-off score. It is then possible to redefine variations in estimated domain scores and domain scores about the cut-off score rather than the mean domain score as is done in classical test theory. Livingston's approach to criterion-referenced test reliability estimation takes the form:

$$K^2(\hat{\pi}, \pi) = \frac{\sigma^2(\hat{\pi}) + (\bar{\hat{\pi}} - \pi_0)^2}{\sigma^2(\pi) + (\bar{\hat{\pi}} - \pi_0)^2} ,$$

where $\hat{\pi}$ is an estimated domain score, π is an examinee's domain score, $\bar{\pi}$ is the mean of the domain scores, $\sigma^2(\hat{\pi})$ is the variance of estimated domain scores about the cut-off score, π_o, and $\sigma^2(\pi)$ is the variance of domain scores about the cut-off score, π_o. It is easy to see that Livingston's estimate of reliability exceeds the classical estimate of reliability given by the expression

$$\sigma^2(\pi)/\sigma^2(\hat{\pi})$$

and increases as $(\bar{\pi}-\pi_o)^2$ increases. In other words, the further the group mean domain score is from the cut-off score, the more reliable the scores are said to be. Notice, even though domain score variance may be zero (a result which would lead to a zero estimate of reliability in classical test theory), it is still possible for Livingston's estimate to exceed zero. Immediately following the publication of Livingston's work there were several published responses to it and replies from Livingston (1972b, 1972c). Harris (1972) made the observation that the standard error of measurement was the same regardless of which approach to reliability was used. This is an important point and is one reason for not rejecting all concepts from classical test theory with criterion-referenced tests. The fact is that the standard error of measurement is one method for setting up confidence bands around domain score estimates (albeit a conservative method). However, this particular point, in and of itself, does not detract from Livingston's formulation or the usefulness of his statistic.

Hambleton and Novick (1973) took issue with Livingston's statement concerning the purpose of criterion-referenced tests. They argued that the difference of an examinee's domain score from a cut-off score was not nearly so important as whether or not an examinee was assigned to the same side of the cut-off score (mastery state) across parallel-form (or retest) administrations of a test. Therefore, they predicted Livingston's approach would have limited usefulness. Of course, this is conjecture on their part, and only time will tell if they are correct. Results reported by Hambleton (1974) do support the Hambleton-Novick position but it is quite possible that others will agree with Livingston.[1]

Shavelson, Block, and Ravitch (1972) took issue with Livingston for reporting the reliability of test scores obtaining by summing across items keyed to different objectives. Shavelson *et al.*, make an important point, (i.e., that reliability information is needed on *each* subset of items measuring an objective included in a test), but it is a point that Livingston can easily handle in his own formulation (Livingston, 1972c). Like, Harris (1972), Shavelson and his colleagues also point out the usefulness of the standard error of measurement of a test. They go on to note that the standard error of measurement is not influenced by Livingston's approach to reliability estimation.

---

[1] Recently we had an opportunity to read an excellent manuscript published in the *Journal of Educational Measurement* by Brennan and Kane (1977). They derive a reliability measure (referred to in their work as an *index of dependability*) for criterion-referenced tests which is developed within the context of generalizability theory (Cronbach, Gleser, Nanda, and Rajaratnam, 1972). Like Livingston, they study examinee domain score deviations from a cut-off score. They point out that there may be occasions when squared-error loss à la Livingston (1972a) is a more appropriate choice of loss function than threshold loss adopted by Hambleton and Novick (1973). They note:

> A squared-error loss function has the advantage of being sensitive to the magnitude of errors, but the disadvantage of being sensitive to all errors of measurement, including those that do not lead to misclassification.
>
> Neither of these loss functions is ideal, and a choice between the two must be made on practical grounds. A threshold loss function is appropriate when there is a sharp cut-off, and all misclassifications are, at least approximately, equal in their impact. A squared-error loss function is likely to be more appropriate when either of these assumptions is violated.

In a follow-up paper which will be published soon, Brennan and Kane (in press) carry on their work with randomly parallel tests and concepts from generalizability theory. The major strength of this new work is that they are able to study many of the approaches to reliability of norm-referenced and criterion-referenced tests within a single framework and thereby draw some important similarities and differences among the approaches.

5.2.2 Reliability of Domain Score Estimates

When there is test score variance it is possible to estimate the standard error of measurement of a criterion-referenced test. (We will assume for convenience, that the test measures only a single objective.) Whereas reliability estimates for a test vary from one sample of examinees to another, the standard error of measurement is generally invariant across samples (Lord and Novick, 1968) and therefore rather useful for interpreting test scores, whether they be scores from a norm-referenced test or a criterion-referenced test. When strictly parallel-tests are available, well-known methods for estimating the standard error of measurement can be used.

In computing the standard error of measurement (SE_M), any of the established procedures for determining the correlation coefficient can be used. That is, repeated measures, parallel forms, or corrected split-half procedures may be used. Further, if one wants to use a lower bound estimate of reliability to compute SE_M, then the Kuder-Richardson formula -21 can be used.

The reader should have two immediate questions: (1) The statistic depends upon a correlation coefficient, so what will be the effect of a restricted range of scores? (2) How do you interpret the statistic?

To answer the first question, we must first understand that as an indicant of error, the smaller the SE_M is in value, the better the test is, re: the more reliable the test is. With this in mind, one next must notice that the formula for the SE_M involves not only the reliability coefficient, but also the standard deviation of test scores. The effect of these two variables is such that, operating in unison, they allow SE_M to be unaffected by the homogeneity of test scores. For example, suppose all the test scores were clumped at the upper end of the test score continuum.

Then r would be low, but $\sqrt{1-r}$ would be a large number. Likewise SD would be low in value, and we can look at their product as being a moderately sized number. If, on the other hand, scores are spread across the continuum, then r is likely to be large in value, as would be SD, but then $\sqrt{1-r}$ would be low, and the result would again be a moderately sized number. The point to be made, using this very simplistic example, is that whereas the reliability coefficient is affected by homogeneity of scores, the standard error of measurement, due to the nature of the formula, is relatively unaffected by the spread of scores in the group of examinees tested. And that is how it should be, for the error (E = X-T) inherent in an examinee's test score, should not depend on the shape of the test score distribution for a group of examinees.

How is the statistic used? What is done, while not technically correct, is as follows: The SE_M, along with an examinee's test score, are used to set up a probability statement about the location of the examinee's (unknown) domain score. An underlying normal distribution is assumed, and hence the area under the normal curve can be used to make some "reasonable" statement about an examinee's domain score. For instance, suppose the SE_M for a test was 5 and a person obtained a score of 50. Based upon the normal curve, 68% of the area lies within one standard deviation to the right and left of the mean. Applied to this example, this could be interpreted, in the non-technical fashion we are using, that the chances are 2 out of 3 that the individual's domain score lies between one standard direction above the mean (50 + 5) and one standard deviation below (50 - 5). Here the test score is used as the mean, and SE_M as the standard deviation. It must be pointed out that on a strictly theoretical level, the above interpretation is wrong on two counts. One, a probability of 2/3 can't really be attached;

the score is either between 45 and 55 (a probability of one) or not (a probability of zero). Two, as mentioned above, the SE_M should be applied to the domain score. So why do it? The answer lies in that, from a practical point of view, we can be reasonably sure about the statement we are making, and this is after all, better than no statement at all.

Example

Suppose an examinee answered 15 out of 20 items correctly on a criterion-referenced test. Suppose also that the test score reliability is .80 and the standard deviation of domain scores is .15.

Questions:

1. What is the value of the standard error of measurement?
2. What is the examinee's domain score estimate?
3. What are the lower and upper limits for an approximately 95% confidence band for the examinee's domain score?

Answers:

1. $SE_M = SD\sqrt{1-r}$

   $= .15 \times .45$

   $= .07$

2. An unbiased domain score estimate for the examinee is .75 (15 divided by 20).
3. Upper limit $= .75 + 2 \times .07$

   $= .89$

   Lower limit $= .75 - 2 \times .07$

   $= .61$

Therefore it can be said that there is an approximately 95% probability that an examinee with a domain score estimate of .75 has a "domain score" somewhere on the interval [.61, .89].

It is often the case that parallel-forms of a criterion-referenced test are constructed by randomly sampling items from a "pool" of test items keyed to an objective. Such tests are referred to as randomly or nominally parallel tests, and typically do not meet the requirements for strictly parallel tests. Randomly parallel tests are examples of the type of measurements for which generalizability theory (Cronbach, et al., 1972) is intended. It is appropriate at this point to turn to generalizability theory to obtain definitions of errors of measurement, of error variance, and formulae for estimating error variance. (See Brennan and Kane [1977, in press] for a more fully developed discussion of the topic.)

Cronbach et al. (1972, p. 25-26), defined three different errors of measurement. One error, Δ_i, is appropriate when the proportion-correct score is taken as an estimate of domain score. The error E_i is appropriate when a linear regression estimate of domain score is made, and the third error δ_i, is appropriate when an estimate of the deviation between the ith examinee's domain score and the mean domain score is made. The second error will not be discussed because typically it is impossible to obtain a regression estimate of domain score on the basis of a single randomly parallel test (see Cronbach et al., 1972, p. 140-146). The third error will not be discussed here because typically there is no reason to estimate the deviation score with criterion-referenced tests.

The error Δ_i is defined as the difference between the observed proportion correct score and the domain score for the ith examinee. Suppose a domain of items exists. Let x_{ij} be the score (0 or 1) for the ith examinee on the jth item. Define $\Delta_{ij} = x_{ij} - \pi_i$.

For an n item test, the error of measurement Δ_i, is $n^{-1} \sum_j \Delta_{ij}$. Cronbach et al., (1972) discussed three variances for Δ_i. These are $\sigma^2_{\Delta|i} = n^{-1} \underset{j}{E} \Delta^2_{ij}$, the error variance for examinee i on an n item test constructed by random sampling of items; $\sigma^2_{\Delta} = \underset{i}{E} \sigma^2_{\Delta|i}$, the average over examinees or $\sigma^2_{\Delta|i}$; and $n^{-2} \underset{i}{E} (\sum_j \Delta_{ij} - \underset{i}{E} \sum_j \Delta_{ij})^2$, the variance over examinees of Δ_i for a given test.

To evaluate the accuracy of a particular test, it would be appropriate to estimate the third variance mentioned above. However, estimation of the quantity requires the administration of several randomly parallel forms, which may not be feasible. Moreover, Cronbach et al. (1972) were pessimistic about the utility of designs for estimating parameters that characterize a given test and so the designs will not be reviewed here. The interested reader is referred to Cronbach et al. (1972, p. 101-102), and references therein for details.

If an n item test has been administered, an estimate of σ^2_{Δ} can be obtained by laying out the item data as a one way ANOVA with examinees as the factor. Item scores are considered to be replications within a level of the examinee factor. The estimate is given by

$$\sigma^2_{\Delta} = \frac{1}{n} MS_{wp} \quad ,$$

where MS_{wp} is the within persons or replications mean square. If several randomly parallel forms of n items each are available then σ^2_Δ can be estimated using the same formula. The proportion-correct scores on the various forms are the replications within a level of the examinee factor.

In principle, it is possible to estimate $\sigma^2_{\Delta|i}$ using the formula

$$\hat{\sigma}^2_{\Delta|i} = \frac{(N-n)}{n^2(n-1)} \sum_{j=1}^{n} (x_{ij} - \hat{\pi}_i)^2$$

for each examinee where $\hat{\pi}_i$ is the observed proportion correct score (estimated domain score) for the ith examinee. The factor (N-n)/n is used when the domain is finite. N is the number of items in the domain. When n is small relative to N, the estimate $\hat{\sigma}^2_{\Delta|i}$ may be quite variable over random samples of items.

Another approach for determining the accuracy of domain score estimates was reported by Millman (1974) and Hambleton, Swaminathan, and Algina (1976). They suggested that the standard error of estimation derived from the binomial test model, given by the expression $\sqrt{\hat{\pi}(1-\hat{\pi})/n}$, could be used to set up confidence bands around domain score estimates. This is a biased estimate and an unbiased estimate is obtained by substituting (n-1) for n in the expression. This is an expression for the standard deviation of errors of measurement for an examinee with domain score π across administrations of n item samples drawn at random from an item pool. A correction $(\frac{N-n}{N})$ can be introduced under the radical sign when the pool of test items is finite. Advantages of this approach are that the estimate of error is a function of domain score, less conservative estimates of error than the one provided by the standard error of measurement are obtained, and the effect of test length on the precision of estimates can be studied easily. In addition, the estimate is relatively easy to compute.

5.2.3 Reliability of Mastery Classification Decisions

Carver (1970) proposed two procedures for assessing the reliability of criterion-referenced tests. The first procedure requires the administration of the same test to two comparable groups, and a comparison of the percentages of examinees that were classified as masters. The second procedure requires the administration of two parallel tests to the same group, and a comparison of the percentage of "masters" on the two tests. With either procedure, the more comparable the percentages, the more reliable the tests are said to be.

Carver (1970) rejected a correlational approach to reliability, arguing that reliability depends on replicability, but replicability does not depend on variance. Carver's procedures were based on the replicability of distributions, while the usual concept of reliability in mental testing is based on the replicability of individual scores. If satisfied, his proposed criteria would provide only the weakest form of evidence for criterion-referenced test reliability; that is, his conditions are necessary but not sufficient to establish test reliability.

Hambleton and Novick (1973) suggested that the reliability of mastery classification decisions should be defined in terms of the consistency of decisions from two administrations of the same test or parallel forms of a test. Suppose examinees are to be classified into m mastery states, the index of reliability tentatively suggested by Hambleton and Novick (1973) was

$$p_o = \sum_{k=1}^{m} p_{kk} \quad ,$$

where p_{kk} is the proportion of examinees classified in the kth mastery state on the two administrations. The index p_o then is the observed proportion of decisions that are in agreement. The p_o statistic has considerable intuitive appeal and is certainly easy to calculate but it suffers from at least one limitation.

Swaminathan, Hambleton and Algina (1974) argued that p_o does not take into account the proportion of agreement that occurs by chance alone and therefore it could give a false impression to users of the extent of mastery classification consistency. They suggested using coefficient κ (Cohen, 1960) as an index of reliability. This coefficient is defined as

$$\kappa = (p_o - p_c)/(1-p_c)$$

where

$$p_c = \sum_{k=1}^{m} p_{k.}\, p_{.k} \quad .$$

The symbols $p_{k.}$ and $p_{.k}$ represent the proportions of examinees assigned to mastery state k on the first and second administrations, respectively. The symbol p_c represents the proportion of agreement that would occur even if the classifications based on the two administrations were statistically independent. Thus, in a sense, it can be argued that κ takes into account the composition of the group, and in this sense, is more group independent than the simple proportion of agreement statistic, p_o.

The properties of κ have been discussed in detail by Cohen (1960, 1968) and Fleiss, Cohen and Everitt (1969) as well as others. For present purposes it is sufficient to note that the upper limit is +1 and can occur only when the marginal proportions for different administrations are equal.

The lower limit is close to -1. The precise lower limit of κ is unimportant in the context of criterion-referenced testing, since any negative value indicates inconsistency and, therefore, unreliable decisions.

The coefficient κ is dependent on all factors that affect the decision-making procedure; the cut-off score, the heterogeneity of the group of examinees, and the method of assigning examinees to mastery states. Millman (personal communication) has suggested that all of these factors be summarized when reporting κ since this information would contribute to its interpretation.

Example

Suppose parallel-forms (denoted Test 1 and 2) of a 4-item test are administered to a group of six students. Suppose further that the cut-off score is set equal to 75%. Consider the data below:

	Test 1					Test 2				
Person	Item 1	Item 2	Item 3	Item 4	Score	Item 1	Item 2	Item 3	Item 4	Score
A	1	0	0	0	1	1	1	0	0	2
B	1	1	1	1	4	1	1	1	0	3
C	1	0	1	0	2	1	1	1	0	3
D	0	0	0	0	0	1	0	0	0	1
E	0	1	1	1	3	1	1	1	0	3
F	0	1	1	0	2	0	1	1	1	3

Question:

What is the value of κ?

Answer:

We require the proportions of examinees who passed and failed on each occasion. The information is reported in the chart below:

Test 2

		Master	Non-master	Marginal Proportion
Test 1	Master	.33	0	.33
	Non-master	.33	.33	.67
	Marginal Proportion	.67	.33	

$$p_o = \sum_{i=1}^{2} p_{ii} = p_{11} + p_{22} = .33 + .33 = .66$$

$$p_c = \sum_{i=1}^{2} p_{i.}\, p_{.i} = p_{1.}\, p_{.1} + p_{2.}\, p_{.2} = (.67)\ (.33) + (.33)\ (.67)$$

$$= .44$$

$$\hat{\kappa} = \frac{p_o - p_c}{1 - p_c}$$

$$\hat{\kappa} = \frac{.66 - .44}{1 - .44}$$

$$= \frac{.22}{.56}$$

$$\hat{\kappa} = .39$$

Of course, in practice one would not estimate κ based on data from only six examinees. The example is offered here for illustrative purposes only.

In criterion-referenced testing situations, it is often the case that administering parallel forms of a test to get an estimate of κ is not feasible. Possible reasons include: (1) The fact that the testing is built into an objectives-based program and the extra testing would take away instructional time, and (2) testing occurs quite often, and two test administrations for each criterion-referenced test would cause the testing process to dominate the students' learning time. Therefore, what is needed is a method of arriving at either κ, or another suitable index, based upon one administration of a test.

The coefficient κ, and p_o, are defined in terms of repeated testings, but it would be very useful to have a procedure for estimating κ or p_o on the basis of a single testing. Such a procedure has been provided by Hunyh (1976a) who prefers κ to p_o. Hunyh (1976a, 1978) assumed that $f(x|\pi)$ is binomial. He also assumed that the marginal distribution of the domain scores is a two parameter beta distribution. From these assumptions it follows that the marginal distribution of test scores obtained by administering any random sample of n items is a negative hypergeometric distribution. Further, the joint distribution of scores obtained by administering two randomly parallel n item tests is a bivariate negative hypergeometric distribution. We will not review his mathematical development here. It is clearly reported in his paper. It is sufficient to say that his solution is workable, although the computations involved in obtaining κ can be tedious when there are a moderate number of possible test scores above the cut-off score. Huynh (1976a) also provided an approximate procedure for estimating κ which appears to work fairly well, if the number of test items is not too small.

Alternative procedures for estimating reliability from a single administration have been provided by Subkoviak (1976) who prefers to work with p_o. While Huynh's approach is more mathematically tractable, it may be far less useful when the number of examinees is small, a fairly common occurrence in objectives-based instructional programs.

Subkoviak (1976) defined a coefficient of agreement for individual i, denoted $p_c^{(i)}$, as the probability of consistent mastery classification of examinee i on parallel forms, denoted X and Y. For the case of two mastery states, this probability is given by

$$p_c^{(i)} = \text{Prob}(X_i \geq c, Y_i \geq c) + \text{Prob}(X_i < c, Y_i < c), \qquad (1)$$

where c is the cut-off score. X_i and Y_i are scores for examinee i on the two tests. The two terms in Equation 1 represent the probability of examinee i being assigned to a mastery state or a non-mastery state on each test administration, respectively. The coefficient of agreement for a group of N examinees is given by

$$p_o = \sum_{i=1}^{N} p_c^{(i)}/N.$$

In order to estimate $p_c^{(i)}$, Subkoviak assumed that for each examinee, scores on the two forms of the criterion-referenced test were independently and identically distributed. Also, he assumed X_i and Y_i for a fixed examinee were identically binomially distributed. This is a questionable assumption although test item responses are usually scores 0 or 1, and item responses are independent. However, the assumption implies that the items making up the test are equally difficult and this will seldom be the case. (Fortunately, Subkoviak addressed this point in his paper and offered a substitute expression -- the compound binomial model -- to handle the more typical case.) With only the two assumptions above, Subkoviak was able to show

$$p(X_i \geq c) = \sum_{x_i=c}^{n} \binom{n}{x_i} \pi_i^{x_i} (1-\pi_i)^{n-x_i} \quad , \qquad (2)$$

and

$$P_c^{(i)} = [p(X_i \geq c)]^2 + [1 - p(X_i \geq c)]^2 \quad . \qquad (3)$$

Once an estimate of an examinee's domain score (π_i), denoted $\hat{\pi}_i$, is obtained, $p(X_i \geq c)$ can be determined by substituting $\hat{\pi}_i$ for π_i in Equation (2). $P_c^{(i)}$ is obtained by substituting the result from Equation (2) into Equation (3). Any of the methods discussed in Unit 8 could be used to estimate an examinee's domain score. Subkoviak suggested in his paper using a regression estimate of π_i, but the merits of this approach would depend on the sample estimates of group mean performance and reliability (as he correctly noted). He also offered several other possible domain score estimates, several of which will be discussed in Unit 8, and others which have been reported by Lord and Novick (1968). A group estimate of the expected proportion of agreement in mastery classifications across parallel-form administrations can be obtained by averaging the values of $p_c^{(i)}$, for $i=1, 2, \ldots, N$, where N is the number of examinees in the group.

It is also possible to obtain an estimate of κ using Subkoviak's method. The only additional information needed is the proportion of examinees assigned to each mastery state on the single test administration. By making the reasonable assumption that these proportions would be the same on a retest or a parallel-form administration, the proportion of agreement expected by chance (p_c) can be obtained by the method introduced earlier ($p_c = p_{1.}\ p_{.1} + p_{2.}\ p_{.2}$). For example, using the first set of test scores from the previous example, it is seen that two of the six examinees would have been assigned to a mastery state based on their test scores. Therefore, $p_{2.} = .33$ and $p_{1.} = .67$, and $p_c = .55$ ($p_c = .33^2 + .67^2$). With a value of p_c and with the p_o estimate from Subkoviak's method, kappa can quickly be calculated, if desired.

Subkoviak's approach to estimating the consistency of mastery classifications across parallel-form administrations can provide either individual or group information, and can be estimated from a single administration of a test. The only two minor problems are that the probability estimates are inflated due to the inclusion of chance agreement, and it is unreasonable to assume all items in a criterion-referenced test are equally difficult. However, on this latter point, Subkoviak has also offered a slightly different model (compound binomial) which is capable of handling the situation.

Subkoviak's method makes it possible to compute the coefficient of agreement in mastery states across occasions for an individual, and also the coefficient of agreement for a group of N persons. Since the formulas developed by Subkoviak are somewhat complex, a step-by-step procedure will be specified and an example will be offered.

The steps in the method are as follows:

1. Obtain an estimate of the proportion of items in the whole domain of items an examinee can answer correctly. A convenient estimate is obtained by setting $\hat{p}_i = \frac{x_i}{n}$,

where $\hat{p}_i$ = proportion-correct score for examinee i,

x_i = his/her test score,

n = total number of items included in the test (measuring the objective of interest).

2. Determine the probability that the examinee's score is greater than or equal to the cutting score (c) using the form of the underlying (binomial) distribution. The probability is given by:

$$P(x_i \geqslant c) = \sum_{x_i=c}^{n} \binom{n}{x_i} \hat{p}_i^{\,x_i} (1-\hat{p}_i)^{n-x_i}$$

where $\hat{p}_i, x_i$, c and n are defined as before, and

$$\binom{n}{x_i} = \frac{n!}{x_i!(n-x_i)!}$$

where $n! = n(n-1)(n-2). . . ,$

[for example:

$4! = 4(3)(2)(1)$].

3. Using the result from step (2), compute the coefficient of agreement for person i using the following formula:

$$p_c^{(i)} = [P(x_i \geq c)]^2 + [1-P(x_i \geq c)]^2$$

4. Finally, compute the coefficient of agreement p_c for a group of N persons, using the following formula:

$$p_c = \frac{\sum_{i=1}^{n} p_c^{(i)}}{N} .$$

The final result, p_c, provides an estimate of the coefficient of agreement for the group had 2 test administrations taken place. The subscript "c" is included to clarify that the coefficient is dependent upon the assigned cut off score. If p_c is high (i.e., close to one), we can be sure that there would be a high degree of consistency of placement into mastery states over the two occasions.

If the number of test items is small, usually a better estimate (than $\hat{p}_i$) of an examinee's domain score can be obtained by using a regression estimate of domain score [$\overset{*}{p}_i = \hat{p}_i r + \bar{p}_i (1-r)$, where r = test reliability, and $\bar{p}_i$ = average proportion-correct score for the examinees], or a Bayesian estimate. (Several promising Bayesian estimates are introduced in Unit 8.) The improved estimate can be substituted for $\hat{p}_i$ in step 2. A quick way to compute r, the test reliability, is to use Kuder-Richardson formula—21 (KR_{21}), given by:

$$KR_{21} = \frac{n}{n-1} \left[1 - \frac{\bar{x}(n-\bar{x})}{nS_x^2} \right]$$

Subkoviak (1976) uses this particular approach. An example is offered next.

Example:

Given the data for test 1 in the previous example, compute p_c. Use regression estimates of domain scores, and KR_{21} as an estimate of test score reliability.

Test 1

Person	Item 1	Item 2	Item 3	Item 4	Score
A	1	0	0	0	1
B	1	1	1	1	4
C	1	0	1	0	2
D	0	0	0	0	0
E	0	1	1	1	3
F	0	1	1	0	2

Reliability Estimate

(a) $\bar{x} = \frac{\Sigma x_i}{N} = \frac{\Sigma x_i}{6} = \frac{12}{6} = 2.0$

(b) $$S_x^2 = \frac{\Sigma(x_i - \bar{x})^2}{N}$$

$$= \frac{(1-2)^2+(4-2)^2+(2-2)^2+(0-2)^2+(3-2)^2+(2-2)^2}{6}$$

$$= \frac{1+4+4+1}{6}$$

$$= 1.67$$

(c) $$r = \frac{n}{n-1} \left[1 - \frac{\bar{x}\ (n-\bar{x})}{nS_x^2}\right]$$

$$= \frac{4}{3} \left[1 - \frac{2(4-2)}{4(1.67)}\right]$$

$$= \frac{4}{3} [1 - .599]$$

$$= .53$$

Regression Estimates of Domain Scores

(d) $\hat{p}_i = r\,(\frac{x_i}{n}) + (1 - r)\,(\frac{\bar{x}}{n})$

$\hat{p}_1 = .53(\frac{1}{4}) + (1 - .53)\,(\frac{2}{4}) = .37$

$\hat{p}_2 = .53(\frac{4}{4}) + (1 - .53)\,(\frac{2}{4}) = .77.$ And, in a like fashion,

$\hat{p}_3 = .50,\ \hat{p}_4 = .24,\ \hat{p}_5 = .63,\ \hat{p}_6 = .50.$

(e) $P(x_i \geqslant c\,) = \sum_{x_i=c}^{4} \binom{4}{x_i} (.37)^{x_i}\ (1 - .37)^{x_i}$

For individual 1,

$$P(x_1 \geqslant 3) = \sum_{x_i=3}^{4} \binom{4}{x_i} (.37)^{x_i} (1 - .37)^{x_i}$$

$$= \binom{4}{3} (.37)^3 (1 - .37)^1 + \binom{4}{4} (.37)^4 (1 - .37)^0$$

and $\binom{4}{3} = \frac{4!}{3!1!} = 4$

$$P(x_1 \geqslant 3) = .2415 + .0187$$

$$= .2602$$

For individual 2, $P(x_2 \geqslant 3) = \sum_{x_i=3}^{4} \binom{4}{x_i} (.74)^{x_i} (.23)^{n-x_i} = .7630$

3, $P(x_3 \geqslant 3) = \sum_{x_i=3}^{4} \binom{4}{x_i} (.5)^{x_i} (.5)^{n-x_i} = .3125$

4, $P(x_4 \geqslant 3) = \sum_{x_i=3}^{4} \binom{4}{x_i} (.24)^{x_i} (.76)^{n-x_i} = .0453$

5, $P(x_5 \geqslant 3) = \sum_{x_i=3}^{4} \binom{4}{x_i} (.63)^{x_i} (.37)^{n-x_i} = .5275$

6, $P(x_6 \geqslant 3) = \sum_{x_i=3}^{4} \binom{4}{x_i} (.5)^{x_i} (.5)^{n-x_i} = .3125$

all computed in a like fashion.

(f) $$p_c^{(i)} = [\underline{P}(x_i \geq c)]^2 + [1 - \underline{P}(x_i \geq c)]^2$$

For individual 1,

$$\begin{aligned} p_c^{(1)} &= [\underline{P}(x_1 \geq 3)]^2 + [1 - \underline{P}(x_1 \geq 3)]^2 \\ &= (.260)^2 + (1 - .260)^2 \\ &= .0676 + .5776 \\ &= .6452 \end{aligned}$$

In a like fashion

$$\begin{aligned} p_c^{(2)} &= (.763)^2 + (1 - .763)^2 &&= .6383 \\ p_c^{(3)} &= (.3125)^2 + (1 - .3125)^2 &&= .5704 \\ p_c^{(4)} &= (.0453)^2 + (1 - .0453)^2 &&= .9134 \\ p_c^{(5)} &= (.5275)^2 + (1 - .5275)^2 &&= .5016 \\ p_c^{(6)} &= (.3125)^2 + (1 - .3125)^2 &&= .5704 \end{aligned}$$

(g) Finally

$$p_c = \frac{\sum_{i=1}^{N} p_c^{(i)}}{N}$$

becomes

$$p_c = \frac{\sum_{i=1}^{6} p_c^{(i)}}{6}$$

and $p_c = .64$.

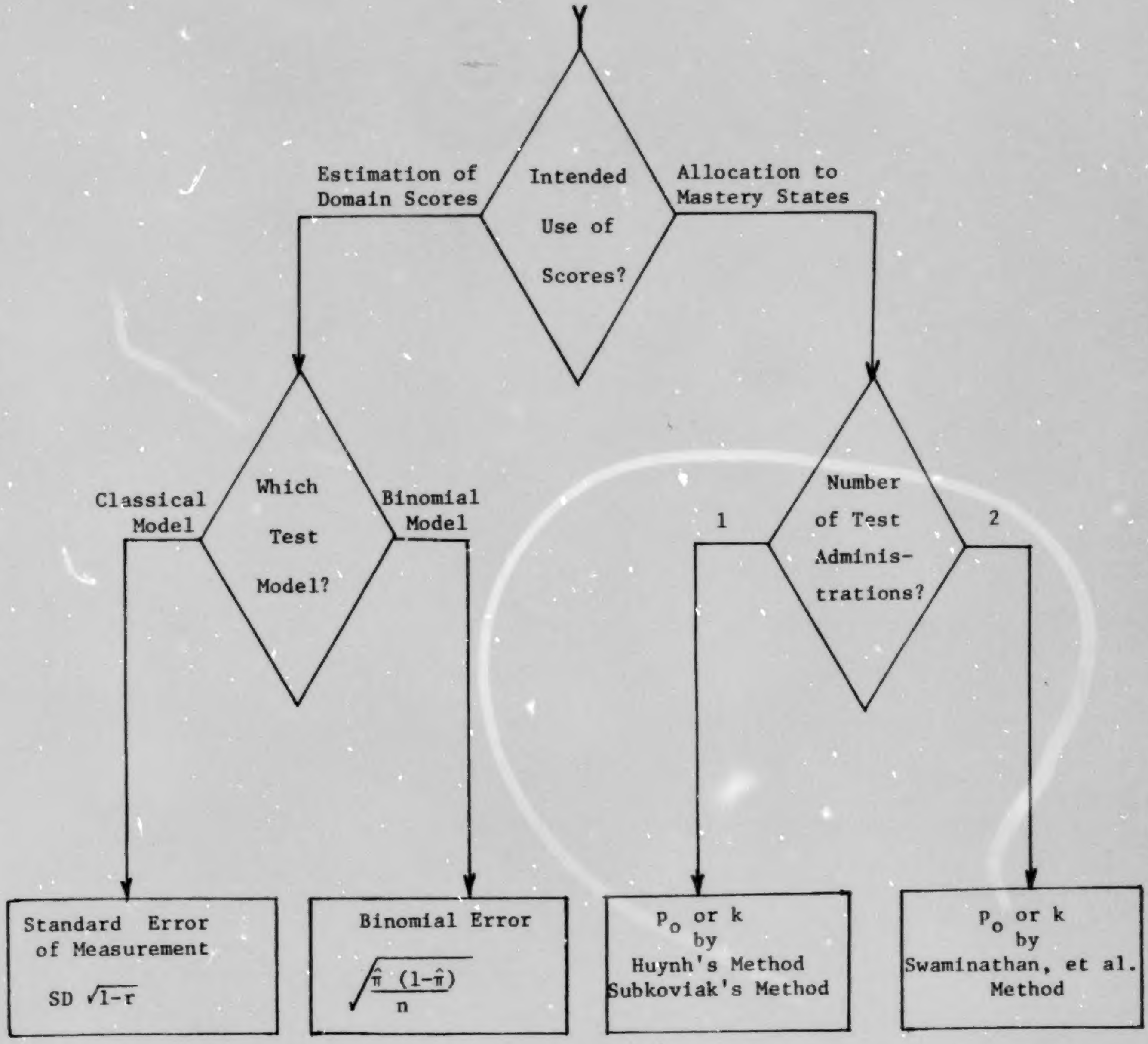
The coefficient of agreement index of .64 obtained using Subkoviak's method should be fairly close in value to the observed proportion of agreement, p_o, which is used in computing κ. That value was .66, and as the reader can see, the values closely coincide.

5.2.4 Summary of the Reliability Discussion

The definition of criterion-referenced reliability chosen should depend upon how the test scores are used. Once a decision has been made about test usage, there are still a number of ways of assessing reliability, depending either upon the underlying distributional assumptions you choose to make or the number of test administrations possible. The chart on the next page will be helpful in summarizing the material in section 5.2.

Perhaps it should be stressed at this point that reliability information (whether one is discussing domain scores or mastery classification decisions) needs to be reported on an objective by objective basis (Hambleton and Novick, 1973; Swaminathan et al., 1975). If a criterion-referenced test measures more than a single objective, as will usually be the case, the test items should be arranged into clusters according to the objectives being measured. Within each of these clusters of items, domain scores may be estimated or mastery classifications made. Whatever the use of the scores, appropriate reliability information should be reported on each use of the scores derived from the test.

Figure 5.2.4 A schematic diagram depicting approaches to reliability assessment.


Intended Use of Scores?
Estimation of Domain Scores
Allocation to Mastery States
Which Test Model?
Classical Model
Binomial Model
Number of Test Adminis- trations?
1
2
Standard Error of Measurement
SD $\sqrt{1-r}$
Binomial Error
$\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$
p_o or k by Huynh's Method Subkoviak's Method
p_o or k by Swaminathan, et al. Method

5.3 Validity of Criterion-Referenced Tests

5.3.1 Introduction

While a number of topics addressed in these instructional materials have been intensely studied by researchers in the criterion-referenced field (for instance, observe the amount of work, albeit disjoint, that has been done on cut-off scores), criterion-referenced test validity is not a member of this highly-researched group. To date, it has remained a minimally explored topic, which is surprising because of its importance. The usefulness of any of the applications of criterion-referenced tests one could name, e.g., to monitor individuals through objectives-based instructional programs, to diagnose learning deficiencies, to evaluate educational and social action programs, to assess competence on certification and licensing examinations, etc., depends directly on the validity of the intended interpretations of the criterion-referenced test scores. Why the lack of validity information? One reason has been offered by Hambleton (1977). He feels that most criterion-referenced test developers simply assume the validity of scores for their intended uses, rather than establishing validity in any formal fashion. Further, the scarce amount of work that has been done has focused on the content validity of criterion-referenced tests, rather than construct validity.

In the sections that follow, we will first clarify the issue of content and construct validity. Then, we will discuss some procedures for establishing content validity, followed by procedures for establishing construct validity.

5.3.2 Clarification of Several Validity Issues

Cronbach (1971), Messick (1975), and Linn (1977) have argued convincingly that to validate interpretations of criterion-referenced test scores (i.e., to determine what is being measured), it is necessary to proceed beyond a consideration of the content validity of a test. Until recently, it was thought that content validity considerations of a criterion-referenced test were sufficient. Messick (1975) states,

> The major problem. . .is that content validity. . . is focused upon test forms rather than test scores, upon instruments rather than measurements. Inferences in educational and psychological measurement are made from scores, and scores are a function of subject responses. Any concept of validity of measurement must include reference to empirical consistency. Content coverage is an important consideration in test construction and interpretation, to be sure, but in itself it does not provide validity. Call it 'content relevance,' if you will, or 'content representativeness,' but don't call it 'content validity' because it doesn't provide evidence for the interpretation of responses or scores.

Content validity is a test characteristic. It will not vary across different groups of examinees or vary over time. However, validity of test score interpretations will vary from one situation to another. For example, if a criterion-referenced test is administered, perhaps by mistake, under highly speeded testing conditions, the validity of interpretations based on test scores obtained from the test administration will be lower than if the test had been administered with more suitable time limits. The content validity of the test does not change, but the validity of any interpretation of the scores does (or at least, can) change from one testing situation to another. Clearly then, content validity evidence is not sufficient to establish "validity of test

300

score interpretations." We must consider construct validity to determine the meaning of a set of scores. To answer the question will require empirical analyses of test scores. In fairness to those who have stressed the importance of content validity for criterion-referenced tests at the expense of other types of validity, we should note that there is a semantic problem. Although the content validity term was being used by many workers in the criterion-referenced testing field, the intended meaning for many was broader than the usual definition of content validity. Thus, many kinds of empirical studies were being done by researchers under the heading of content validity. For example, Rovinelli and Hambleton (1977) discuss both the use of content specialists' ratings and empirical data under the heading of content validity. It would have been helpful to label the studies for what they are, i.e., construct validation studies.

Perhaps at the risk of belaboring the point, we will repeat a point which is stated frequently but appears to have been neglected with criterion-referenced tests. It is that the concept of "validity of measurement" refers to the scores and not to the test. Messick (1975) notes,

> One validates, not a test, but an interpretation of data arising from a specific procedure.

Linn (1977) also makes the same point,

> Questions of validity are questions of the soundness of the interpretation of a measure. Thus, it is the interpretation rather than the measure that is validated. Measurement results may have many interpretations which differ in their degree of validity and in the type of evidence required for the validation process.

The resolution of the validity question would seem to be this: A content validity study is essential at the test development stage, and the content validity of a criterion-referenced test will influence the kind of test score interpretations that are possible. Also, it is most important to conduct construct validation studies to validate the intended use of the test scores. The nature of the construct validation studies will depend on the intended use of the test scores.

In spite of its stated importance, it cannot be argued either that the nature of content validation studies with criterion-referenced tests is well-understood. Guion (1977), for one, discusses many of the problems surrounding the topic. It is only recently that any progress (i.e., the development and field testing of content validation methods with criterion-referenced tests) has been made (Millman, 1974; Popham, 1978; Rovinelli and Hambleton, 1977).

In summary, content validation studies will address the matter of content relevance of material that finds its way into a test. Construct validation studies will relate to the matter of "meaning of scores." These studies will include correlational, experimental, as well as other methods of investigation.

Contributing to some confusion among test developers is the proliferation of new validity terms. Domain validity, descriptive validity, functional validity, domain-selection validity, and incremental validity are but five. Our preference is to stay with the standard terms, but define them clearly.

5.3.3 Content Validation Studies

Content validity has proven to be a fuzzy and confusing concept, even for norm-referenced test developers. Some like Guion (1977) prefer the term "content representativeness" to "content validity" because the former expression is more descriptive. For criterion-referenced test developers, content validity refers to the matter of how well an observed sample of behaviors reflects the larger domain of behaviors included in a domain specification written to define an objective. "Content" is broadly defined to include material from the cognitive, affective, and psychomotor domains.

Generally speaking, the quality of criterion-referenced test items can be determined by the extent to which they reflect, in terms of their content, the domains from which they were derived. The problem here is one of item validation; unless one can say with a high degree of confidence that the items in a criterion-referenced test measure the intended objectives, any use of the test score information is questionable. When domain specifications are utilized, the domain definition is never really precise enough to assume a priori that the items are valid. Thus the quality of the items must be determined in a context independent from the process by which the items were generated. This is an a posteriori approach to item validation. Some procedures have been designed to assess whether or not a direct relationship between an item and a domain or objective exists through analysis of data collected after the item is written (Hambleton & Fitzpatrick, in preparation; Popham, 1978).

There are two approaches which may be used to establish the (content) validity of test items. The first approach, and the approach we feel holds the most merit, involves the judgment of test items by content specialists. The judgments that are made concern the extent of "match" between the test items and the domain they are designed to measure. Questions asked of content specialists about content validity of test items can be reduced to two important ones:

1. Is the format and content of an item appropriate to measure some part of the domain specification?
2. Does the available set of test items adequately sample a particular domain?

A second approach is to apply empirical techniques to examinee response data in much the same way empirical techniques are applied in norm-referenced test development. In fact, along with some recently developed empirical procedures for criterion-referenced tests, several norm-referenced test items statistics can (and should) be used. The problem is to ensure that these statistics are used and interpreted correctly in the context of criterion-referenced test development. Item statistics should be used to detect aberrant items that need to be reworked, and not to make final decisions about which items are to be included in a test. An excellent review of item statistics for use with criterion-referenced tests has been prepared by Berk (1978).

The first question is studied by comparing test items generated by different content specialists and analyzing the judgments of content specialists about items relative to the domain they were developed to measure. The second question is a difficult one to investigate, unless the domain of items is completely specified. The question can be investigated by Cronbach's (1971) interesting but somewhat impractical duplication experiment.

5.3.4 Construct Validation Studies

Content validity evidence does not address the matter of validity of criterion-referenced test score interpretations. Content validity is a characteristic of the test. Clearly it is essential to establish validity of score interpretations and therefore construction validation studies are needed. We like Messick's (1975) definition of construct validation:

> Construct validation is the process of marshalling evidence in the form of theoretically relevant empirical relations to support the inference that an observed response consistency has a particular meaning (p. 955).

Messick (1975) offers several explanations for why construct validation studies have not been more common in educational measurement. For one, content validity of criterion-referenced tests was seen as sufficient. Second, criterion-referenced test score distributions are often homogeneous (for example, it often happens that before instruction most individuals do poorly on a test, and after instruction, most individuals do well). Correlational methods do not work very well with homogeneous distributions of scores because of score range restrictions. But, as Messick (1975) has noted,

> . . . construct validation is by no means limited to correlation coefficients, even though it may seem that way from the prevalence of correlation matrices, internal consistency indices, and factor analysis (p. 958).

Construct validation studies should begin with a definite statement of the proposed interpretation. This will provide direction for the kind of evidence that is worth collecting. Cronbach (1971, p. 483) notes, "Investigations to be used for construct validation, then, should be purposeful rather than haphazard." Later, when all of the data is collected and analyzed, a final conclusion as to the validity of the intended interpretation can be offered.

Let us next review some of the investigations that could be conducted.

Guttman Scalogram Analysis

It frequently occurs that objectives can be arranged linearly or hierarchically. Guttman scaling is a relevant procedure for the construct validation of criterion-referenced test items in situations where the objectives can be organized into either a linear or hierarchical sequence. To use Guttman's scalogram analysis as a technique in an item validation methodology, one would first need to specify the hierarchical structure of a set of objectives. To the extent that examinee responses to the test items intended to measure objectives in the hierarchy are predictable from a knowledge of the hierarchy, one would have evidence to support the construct validity of the test items as measures of the intended objectives. On the other hand, we should note that in situations where examinee item responses are not predictable, one of three situations has occurred:

1. The hierarchy is incorrectly specified;

2. The items are not valid measures of the intended objectives; or
3. A combination of the two explanations.

More precise specifications for the utilization of Guttman scaling will of course be needed before the method can be fully implemented in the validation process for criterion-referenced test items.

Factor Analysis

While factor analysis is a commonly employed procedure for the dimensional analysis of items in a norm-referenced test, or of scores derived from different norm-referenced tests, it has rarely, if ever, been used in construct validation studies of criterion-referenced test scores. Perhaps one reason for its lack of use is that the usual input for factor analytic studies is correlations, and correlations are often low between items on a criterion-referenced test, or between criterion-referenced test scores and other variables because score variability is often not very great. However, the problem can be remedied by choosing a sample of examinees with a wide range of ability. Required is a group of masters and non-masters. The research problem in the language of factor analysis, becomes a problem of determining whether or not the factor pattern matrix has a prescribed form. One would expect to obtain as many factors in a factor solution as there are objectives covered in a test, and with items "loading" on only the factor (or objective) that they were designed to measure. Items deviating from this pattern could be carefully studied for flaws.

Similarly, scores from many criterion-referenced tests could be factor analyzed and the resulting structure could be compared to some structure specifying a theoretical relationship among the tests.

In addition, scores from other tests might be correlated to provide a base for convergent and divergent studies.

Experimental Studies of Sources of Invalidity

There are many sources of error that can reduce the validity of an intended interpretation of a test score. Suppose, for example, we estimated an examinee to have an 80% level of performance on a test measuring "ability to identify the main idea in paragraphs." Is 80% an accurate assessment of the examinee's ability? We might ask about the influence of many factors:

1. How clear were the test directions?
2. Was there any confusion in using the answer sheets?
3. Was the test administered under speeded testing conditions?
4. Was the examinee motivated?
5. Was the examinee interested in the content of the paragraph?
6. Was the vocabulary suitable?
7. What role did test-taking skills play in the examinee's performance?
8. Was the item format suitable for measuring the desired skill?
9. At what time during the day was the test administered?
10. Were the physical surroundings suitable?

To the extent that any of these (and many other) factors influence test scores, the usefulness of the test scores is reduced.

Required are experimental studies of potential sources of error to determine their effect on test scores. Results of these studies can be used to further clarify domain specifications. For example, if we discovered item format influenced test scores, we could include in the domain specifications which item type should be used, after we determined which produced the most construct valid test scores.

5.3.5 Summary

In this section of the instructional materials on criterion-referenced tests, we have tried to clarify the present status of criterion-referenced test validity. We have explicated the differences between content and construct validity and have discussed procedures for establishing both content and construct validity.

5.4 Norms for Interpreting Criterion-Referenced Test Scores

A question occasionally asked by criterion-referenced test users is whether or not it makes sense to use norms with such tests. Will the use of norms data enhance or erode the interpretability of a criterion-referenced test? Popham (1976) has discussed this point, and the discussion which follows contains many of his ideas.

For norm-referenced tests, where there is only a general specification of the content area being addressed, test scores derive their meanings through comparisons to norm group data. For criterion-referenced tests, where clearly described domains of behaviors are specified, test scores can derive their meaning by being referenced to this domain of behaviors. The major difference is, then, in the ability of these two types of tests to describe exactly what a student can do. In norm-referenced testing, where there is only a general description of the content area being measured, little can be said about what an individual can do. In criterion-referenced testing, where the content area is clearly defined, absolute statements about what an individual can do are possible. The problem is that while an accurate description of what an individual can do is useful, it is often not enough.

A decision-maker often wants to know more than what a student can do; he/she wants to evaluate the observed level of performance. The test performance of suitable norm groups provides an excellent basis upon which to gain additional insights into what should constitute an acceptable level of test performance. For example, if it is known how well a group of students performed in a program the previous year, or how well students in a neighboring school district performed on a particular test, it becomes possible to provide a framework for viewing and interpreting new individual and group performance on the same test.

An expressed fear about using norms data with criterion-referenced tests is that through the use of such data, criterion-referenced testing procedures will somehow be forsaken for norm-referenced ones. In other words, the fear is that by adding norms, the procedures and interpretability of criterion-referenced tests will be eroded. This is not so; use of norms will not do anything to the descriptive quality of the test. Rather the use of norms will supplement the basic interpretations. Test scores will then be able to be interpreted in an absolute fashion in reference to the objectives and in a comparative fashion in reference to the norms data. The best one can hope for with a norm-referenced test is test score interpretation on a comparative basis. Hence, the fear of eroding the basic nature of a criterion-referenced test by introducing norms is unfounded.

One fear that is founded, according to Popham (1976), is that:

> . . .users of criterion-referenced tests will
> unthinkingly rely on normative data as a
> determiner of performance standards.

Rather than using the norms data as a sole determiner of standards, a number of other viable procedures can be used. The reader should refer to Unit 6 and the discussion of cut-off scores.

Given that the use of norms data can supplement the interpretability of criterion-referenced test scores, how then should a test developer go about the task of collecting and representing the norms data? Little that is new can be said about the collection of the data; the test should be administered to a representative sample of students from the norm group (or groups) of interest. Of the numerous ways of presenting norms data, one possibility is percentile ranks. The percentile rank a student receives is defined as the percentage of students in the norm or reference group who score equal to or below the student's test score. What follows is a brief discussion of how to obtain percentile ranks. The reader should consult any of the standard test and measurement texts listed in the reference section of Unit 2 for a more in-depth discussion of percentile ranks.

One popular method for calculating percentile ranks is described next and a simple example is offered.

Computation of Percentile Ranks

1. Prepare a frequency distribution (f) of the scores.
2. Find the cumulative frequency (CF), the number of persons scoring lower than the score in question, by summing the frequency (f) of scores below the score in question.
3. Find the culultative frequency to the mid-point (CF_{mp}) by adding one-half the number of scores in the interval to CF:

   $$CF_{mp} = CF + .5f_i$$

4. Find the cumulative proportion (CP) by dividing CF_{mp} by N, the total number of scores.
5. Multiply CP x 100.

The following example, based on an N of 25, is offered to demonstrate the steps listed above. Usually a much larger sample size is used in setting up a table of percentile ranks. A single test score has too much influence on the distribution of percentile ranks for small N.

Raw Score	f	CF	CF_{mp}	CP	Percentile Rank
10	2	23	24.5	.98	98
9	3	20	21.5	.86	86
8	7	13	16.5	.66	66
7	6	7	10	.40	40
6	4	3	5	.20	20
5	2	1	2	.08	8
4	1	0	.5	.02	2

5.5 References

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9.

Berk, R. A. Criterion-referenced test item analysis and validation. Paper presented at the First Annual Johns Hopkins University National Symposium on Educational Research, Washington, 1978.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289.

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. Psychometrika, in press.

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluative research: Strategies and methods. Washington: American Institutes for Research, 1970.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. Psychological Bulletin, 1968, 70, 213-220.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), Educational measurement. (2nd ed.) Washington: American Council on Education, 1971.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: John Wiley & Sons, 1972.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. Psychological Bulletin, 1969, 72, 323-327.

Guion, R. M. Content validity: The source of my discontent. Applied Psychological Measurement, 1977, 1, 1-10.

Haladyna, T. M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99.

Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44, 371-400.

Hambleton, R. K. Validation of criterion-referenced test score interpretations. A paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, 1977.

Hambleton, R. K., & Fitzpatrick, A. Review techniques for criterion-referenced test items. Manuscript in preparation.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159-170.

Hambleton, R. K., Swaminathan, H., & Algina, J. Some contributions to the theory and practice of criterion-referenced testing. In D. N. M. de Gruijter, and L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: Wiley, 1976.

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement*, 1972, *9*, 27-29.

Huynh, H. On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement*, 1976, *13*, 253-264. (b)

Huynh, H. Reliability of multiple classification. *Psychometrika*, 1978, *43*, 317-325.

Linn, R. L. Issues of validity in measurement for competency-based programs. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1977.

Livingston, S. A. Criterion-referenced applications of classical test theory. *Journal of Educational Measurement*, 1972, *9*, 13-26. (a)

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability of coefficient for criterion-referenced tests." *Journal of Educational Measurement*, 1972, *9*, 31. (b)

Livingston, S. A. Reply to Shavelson, Block and Ravitch's "Criterion-referenced testing: Comments on reliability." *Journal of Educational Measurement*, 1972, *1*, 139-140. (c)

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Messick, S. A. The standard problem: Meaning and values in measurement and evaluation. *American Psychologist*, 1975, *30*, 955-966.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, California: McCutchan Publishing Co., 1974.

Popham, W. J. Normative data for criterion-referenced tests. *Phi Delta Kappan*, 1976, *58*, 593-594.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.

Rovinelli, R. J., & Hambleton, R. K. On the use of content specialists in the assessment of criterion-referenced test item validity. Tijdschrift voor Onderwijsresearch, 1977, 2, 49-60.

Shavelson, R. J., Block, J. H., & Ravitch, M. M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, 133-137.

Subkoviak, M. Estimating reliability from a single administration of a criterion-referenced test. Journal of Educational Measurement, 1976, 13, 265-275.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-268.

Unit 6

Issues and Methods for Standard-Setting[1]

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

---

[1] Portions of material in this unit are from Hambleton and Eignor (1979) and Eignor (1979) (see references).

Table of Contents

Page

6.0 Overview of the Unit. 1

6.1 Introduction. 2

6.2 Some Issues in Standard Setting 5

  6.2.1 Uses of Cut-Off Scores in Decision-Making. 5

6.3 Distinction Between Continuum and State Models. 13

6.4 Traditional and Normative Procedures. 15

6.5 Consideration of Several Promising Standard Setting Methods 18

6.6 Judgmental Methods. 20

  6.6.1 Item Content 20
  6.6.2 Guessing and Item Sampling 30

6.7 Empirical Methods 31

  6.7.1 Data From Two Groups 31
  6.7.2 Decision-Theoretic Procedures. 38
  6.7.3 Empirical Methods Depending Upon a Criterion Measure 43
  6.7.4 Educational Consequences 48

6.8 Combination Methods 52

  6.8.1 Judgmental-Empirical 52
  6.8.2 Bayesian Procedures. 55

6.9 Some Procedural Steps in Standard Setting 57

  6.9.1 Preliminary Considerations 58
  6.9.2 Classroom Testing. 59
  6.9.3 Basic Skills Testing for Annual Promotion and High School Graduation. 64
  6.9.4 Professional Licensing/Certification Testing . . . 68

6.10 Summary 70

6.11 References. 72

6.0 Overview of the Unit

In this Unit, some of the issues involved in standard setting along with methods for standard-setting are reviewed. The review will draw on the work of Millman, Meskauskas, and Glass and incorporate many of the newer standard-setting methods. The standard-setting methods are organized into three categories, judgmental methods, empirical methods, and combinations of judgment and empirical methods. Procedures for setting standards to accomplish three primary uses of criterion-referenced testing are discussed in a final section of the paper.

6.1 Introduction

In a recent review of the criterion-referenced testing field, Hambleton, Swaminathan, Algina, and Coulson (1978) delineated two major uses for test scores derived from criterion-referenced tests: domain score estimation and the allocation of examinees to mastery states. The second use, the allocation of examinees to mastery states, requires the setting of a performance standard, or cut-off score.

Based upon an individual's score on a test, where the test is a representative sample of the subject domain, a mastery/non-mastery decision concerning the domain from which the item sample was drawn is sought. Millman (1973) summarizes the situation well:

> Of interest is the proportion of such items a student can pass. It is assumed that some educational decision, e.g., the nature of subsequent instruction for the student, is conditional upon whether or not he exceeds a proficiency standard when administered a sample of items from the domain. Thus, attention is directed toward the individual examinee and his performance relative to the standard rather than toward producing indicators of group performance.

Thus, it can be seen that in this criterion-referenced testing situation, a cut-off score (there can be multiple cut-off scores on the domain score scale although usually only one is set) must be set, in order to make a decision about an individual's mastery status. The results of this decision will depend upon the context within which the test is being used. As an example, consider the Mastery Learning paradigm (Block, 1972). In this situation, if a student's score exceeds the cutting score, he/she is advanced to the next unit of instruction. If the student's score falls below the standard, remedial activities are prescribed. It is important to understand that the decision being made is on the level of the individual, and as such, the status of other individuals does not enter into the decision. As a second example, consider the use of criterion-referenced tests to provide test data relative to a set of basic skills which students must demonstrate mastery of (i.e., achieve specified levels of performance) in order to graduate from high school. In this context, decisions are very important because whether or not students can graduate will depend on their criterion-referenced test score performance and the resulting master/non-mastery decisions which are made.

These situations can be contrasted with the setting of standards for norm-referenced tests, which is considerably less complex. Since for tests constructed to yield norm-referenced interpretations, an individual is compared to others, it makes sense to set a passing or cut-off score so that a certain percent of the students pass. If, for instance, only 20% of the students taking an exam can be placed in an enrichment program, then a passing score that passes 20% of the students would make sense.

Given what has just been said about the importance of cut-off scores for proper criterion-referenced test score usage, one would think that this would be well-researched and documented area. This is simply not the case. Most of the work done to date has been concerned with the suggestion of possible methods, perhaps twenty-five in number, rather than with actual empirical investigations. In addition to the individual work done, there have been two excellent reviews of cut-score procedures advanced (Millman, 1973; Meskauskas, 1976), and one recent review that was highly critical of the field (Glass, 1978a, 1978b).

6.2 Some Issues in Standard Setting

One of the primary purposes of criterion-referenced testing is to provide data for decision-making. Sometimes the decisions are made by classroom teachers concerning the monitoring of student progress through a curriculum. On other occasions, promotion, certification and/or graduation decisions are made by school, district, and state administrators.

Glass (1978a) was rather critical of measurement specialists for giving too little attention to the problem of determining cut-off scores [he notes, "A common expression of wishful thinking is to base a grand scheme on a fundamental, unsolved problem." (p. 1)]. On the other hand, a considerable amount of criterion-referenced testing research has been done. Not all uses of criterion-referenced tests require cut-off scores (for example, description), and moreover, the problem does not really arise until a criterion-referenced test has been constructed. Also, it should not be forgotten that problems associated with cut-off scores are difficult and so solutions are going to require more time.

6.2.1 Uses of Cut-off Scores in Decision Making

A "cut-off score" is a point on a test score scale that is used to "sort" examinees into two categories which reflect different levels of proficiency relative to a particular objective measured by a test. It is common to assign labels such as "masters" and "non-masters" to examinees assigned to the two categories. It is not unusual either to assign examinees to more than two categories based on their test performance (i.e., sometimes multiple cut-off scores are used) or to use cut-off scores that vary from one objective to another (this may be done when it is felt that a set of objectives differ in their importance).

It is important at this point to separate three types of standards or cut-off scores. Consider the following statement:

> School district A has set the following target—
> It desires to have 85% or more of its students
> in the second grade achieve 90% of the reading
> objectives at a standard of performance equal
> to or better than 80%.

Three types of standards are involved in the example:

1. The 80% standard is used to interpret examinee performance on each of the objectives measured by a test.
2. The 90% standard is used to interpret examinee performance across all of the objectives measured by a test.
3. The 85% standard is applied to the performance of second graders on the set of objectives measured by a test.

In this unit, only the first use of standards or cut-off scores will be considered.

In what follows it is important to separate the theoretical arguments for or against the uses of cut-off scores from the uses and misuses of cut-off scores in practical settings. For example, it is well-known that cut-off scores are often "pulled from the air" or set to (say) 85% because that is the value another school district is using. But, the fact that cut-off scores are being determined in a highly inappropriate way is obviously not grounds for rejecting the concept of a "cut-off score." If the concept is appropriate for some particular use of a criterion-referenced test, the task becomes one of training people to set and to use cut-off scores properly (Hambleton, 1978).

Four questions with respect to the use of cut-off scores with criterion-referenced tests require answers:

1. Why are cut-off scores needed?
2. What methods are available for setting cut-off scores?
3. How should a method be selected?
4. What guidelines are available for applying particular methods successfully?

1. Why are cut-off scores needed?

An answer to the question depends on the intended use (or uses) of the test score information. Consider first objectives or competency-based programs since it is with these types of programs that criterion-referenced tests and cut-off scores are most often used. Objectives-based programs, in theory are designed to improve the quality of instruction by (1) defining the curricula in terms of objectives, (2) relating instruction and assessment closely to the objectives, (3) making it possible for individualization of instruction, and (4) providing for on-going evaluation. Hard evidence on the success of objectives-based programs (or most new programs) is in short supply but there is some evidence to suggest that when objectives-based programs are implemented fully and properly they are better than more "traditionally-oriented" curricula (Klausmeier, Rossmiller, & Saily, 1977; Torshen, 1977). Individualization of instruction is "keyed" to descriptive information provided by criterion-referenced tests relative to examinee performance on test items measuring objectives in the curriculum. But descriptive information such as "examinee A has answered correctly

85% of the test items measuring a particular objective" must be evaluated and decisions made based upon that interpretation. Has a student demonstrated a sufficiently high level of performance on an objective to lead to a prediction that she/he has a good chance of success on the next objective in a sequence? Does a student's performance level indicate that he/she may need some remedial work? Is the student's performance level high enough to meet the target for the objective defined by teachers of the curriculum? In order to answer these and many other questions it is necessary to set standards or cut-off scores. How else can decisions be made? Comparative statements about students (for example, Student A performed better than 60% of her classmates) are largely irrelevant. Carefully developed cut-off scores by qualified teams of experts can contribute substantially to the success of an objectives-based program (competency-based program or basic skills program) because cut-off scores provide a basis for effective decision-making.

There has also been criticism (Glass, 1978a) of the use of cut-off scores with "life skills" or "survival skills" tests. These are terms currently popular with State Departments of Education, School Districts, Test Publishers, and the press. Of course, Glass is correct when he notes that it would be next to impossible to validate the classifications of examinees into "mastery states", i.e., those predicted to be "successful" or "unsuccessful" in life. On the other hand, if what is really meant by the term "life skills" (say) is "graduation requirements," then standards of performance for "basic skills" or "high school competency" tests can probably be set by appropriately chosen groups of individuals (Millman, personal communication).

2. If cut-off scores are needed, what methods are available for setting them?

Numerous researchers have catalogued many of the available methods (Hambleton & Eignor, 1979; Hambleton et al., 1978; Jaeger, 1976; Millman, 1973; Meskauskas, 1976; Shepard, 1976). Many of these methods have also been reviewed by Glass (1978a). It suffices to say here that there exist methods based on a consideration of (1) item content, (2) guessing and item sampling, (3) empirical data from mastery and non-mastery groups, (4) decision-theoretic procedures, (5) external criterion measures, and (6) educational consequences. These methods will be considered in detail in sections 6.6, 6.7, and 6.8.

What is clear is that all of the methods are arbitrary and this point has been made or implied by everyone whose work we have had an opportunity to read. The point is not disputed by anyone we are aware of. But as Glass (1978a) notes, "arbitrariness is no bogeyman, and one ought not to shrink from a necessary task because it involves arbitrary decisions" (p. 42). Popham (1978) has given an excellent answer to the concern expressed by some researchers about arbitrary standards:

> Unable to avoid reliance on human judgment as the chief ingredient in standard-setting, some individuals have thrown up their hands in dismay and cast aside all efforts to set performance standards as arbitrary, hence unacceptable.
>
> But Webster's Dictionary offers us two definitions of arbitrary. The first of these is positive, describing arbitrary as an adjective reflecting choice or discretion, that is, "determinable by a judge or tribunal." The second definition, pejorative in nature, describes arbitrary as an adjective denoting capriciousness, that is, "selected at random and without reason."

> In my estimate, when people start knocking the standard-setting game as arbitrary, they are clearly employing Webster's second, negatively loaded definition.
>
> But the first definition is more accurately reflective of serious standard-setting efforts. They represent genuine attempts to do a good job in deciding what kinds of standards we ought to employ. That they are judgmental is inescapable. But to malign all judgmental operations as capricious is absurd. (p. 168)

And, in fact, much of what we do is arbitrary in the positive sense of the word. We set fire standards, health standards, environmental standards, highway safety standards, (even standards for the operation of nuclear reactors), and so on. And in educational settings, it is clear that teachers make arbitrary decisions about what to teach in their courses, how to teach their material, and at what pace they should teach. Surely, if teachers are deemed qualified to make these other important decisions, they are equally qualified to set standards or cut-off scores for the monitoring of student progress in their courses. But what if a cut-off score is set too high (or low) or students are misclassified? Through experience with a curriculum, with high quality criterion-referenced tests, and with careful evaluation work, standards that are not "in line" with others can be identified and revised. And for students who are misclassified, there are some redeeming features. Those that perform below the standard will be assigned remedial work and the fact that they performed below the cut-off score suggests that they could not be too far above it (this would be true for most of the students about whom false-negative errors are made) and so the review period will not be a total waste of time.

And for those students who are misclassified because they scored above a cut-off score, they will be tested again. It is possible the next time the error will be caught (particularly if the objectives are sequential). A comment by Ebel (1978) is particularly appropriate at this point:

> Pass-fail decisions on a person's achievement in learning trouble some measurement specialists a great deal. They know about errors of measurement. They know that some who barely pass do so only with the help of errors of measurement. They know that some who fail do so only with the hindrance of errors of measurement. For these, passing or failing does not depend on achievement at all. It depends only on luck. That seems unfair, and indeed it is. But, as any measurement specialist can explain, it is also entirely unavoidable. Make a better test and we reduce the number who will be passed or failed by error. But the number can never be reduced to zero. (p. 549)

The consequences of false-positive and false-negative errors with basic skills assessment or high school certification tests are however considerably more serious and so more attention must be given to the design of these testing programs (for example, content covered by the tests, the timing of tests, and decisions made with the test results). Considerably more effort must also be given to test development, content validation, and setting of standards.

3. How should a method be selected?

There are many factors to consider in selecting a method to determine cut-off scores. For example,

1. How important are the decisions?
2. How much time is available?
3. What resources are available to do the job?
4. How capable are the appropriate individuals of applying a particular method successfully?

The most interesting workwe have seen to date regarding the selection of a method was offered by Jaeger (1976). He considers several methods for determining cut-off scores, several approaches for assigning examinees to mastery states, and various threats to the validity of assignments. While Jaeger's work is theoretic, it provides an excellent starting point for anyone interested in initiating research on the merits of different methods. One thing seems clear from his work—all of the methods he studied appear to have numerous potential drawbac[illegible] and so the selection of a method in a given situation should be made carefully.

4. What guidelines are available for applying particular methods successfully?

Unfortunately, there are relatively few sets of guidelines available for applying any of the methods. In our judgment, Zieky and Livingston (1977) have provided a very helpful set of guidelines for applying several methods (the popular Nedelsky method and the Angoff method are two of the methods included). Some new work by Popham (1978) is also very helpful. More materials of this type and quality are needed. Some procedural steps for standard-setting with respect to three important uses of tests — (1) daily classroom assessment, (2) basic skills assessment for yearly promotions and high school certification, and (3) professional licensing and certification are provided in section 6.9.

6.3 Distinction Between Continuum and State Models

The basic difference between continuum and state models has to do with the underlying assumption made about ability. According to Meskauskas, two characteristics of continuum models are:

1. Mastery is viewed as a continuously distributed ability or set of abilities.
2. An area is identified at the upper end of this continuum, and if an individual equals or exceeds the lower bound of this area, he/she is termed a master.

State models, rather than being based on a continuum of mastery, view mastery as an all-or-none proposition (i.e., either you can do something or you cannot). Three characteristics of state models are:

1. Test true-score performance is viewed as an all-or-nothing state.
2. The standard is set at 100%.
3. After a consideration of measurement errors, standards are often set at values less than 100%.

There are at least three methods for setting standards that are built on a state model conceptualization of mastery. The models take into account measurement error, deficiencies of the examination, etc., in "tempering" the standard from 100%. These methods have been referred to by Glass (1978a) in his review of methods for setting standards as "counting backwards from 100%." State model methods advanced to date include the mastery testing evaluation model of Emrick (1971), the true-score model of Roudabush (1974), and some recently advanced statistical models of Macready and Dayton (1977). However, since state models are somewhat less usefulness than continuum models in elementary and secondary school testing programs, they will not be considered further here. Our failure to consider them further however, should not be interpreted as a criticism of this general approach to standard-setting. The approach seems to be especially applicable with many performance tests (Hambleton & Simon, in preparation).

6.4 Traditional and Normative Procedures

Before discussing the various continuum models of standard setting, two other models for standard-setting should be mentioned. These methods, which seem to have limited value in setting standards, have been referred to by a variety of names. We will call them "traditional standards" and "normative standards."

Traditional standards are standards that have gained acceptance because of their frequent use. Classroom examples include the decision that 90 to 100 percent is an A, 80 to 89 percent is a B, etc. It appears that such methods have been used occasionally in setting standards.

"Normative" standards refer to any of three different uses of normative data, two of which are, at best, questionable. In the first method, use is made of the normative performance of some external "criterion" group. As an example, Jaeger (1978) cites the use of the Adult Performance Level (APL) tests by Palm Beach County, Florida schools. Test performance of groups of "successful" adults were used to set standards for high school students. Such a procedure can be criticized on a number of grounds. Jaeger (1978) points out that society changes, and that standards should also change. Standards based on adult performance may not be relevant to high school students. Shepard (1976) points out that any normatively-determined standard will immediately result in a multitude of counterexamples. Further, Burton (1978) suggests that relationships between skills in school subjects and later success in life are not readily determinable, hence, observing the degree of achievement on the test of some "successful" norm group makes little sense. Jaeger (1978) goes on to say: "There are no empirically tenable survival standards on school-based skills that can be justified through external means."

A second way of proceeding with normative data is to make a decision about a standard based solely on the distribution of scores of examinees who take the test. Such a procedure circumvents the "minimum test score for success in life" problem, but the procedure is still not useful for setting standards. For example, Glass (1978a) cites the California High School Proficiency Examination, where the 50th percentile of graduating seniors served as the standard. What can be said of a procedure where whether or not an individual passes or fails a minimum competency test depends upon the other individuals taking the test? In the California situation, the standard was set with no reference at all to the content of the test or the difficulty of the test items.

The third use of normative data discussed in the literature concerns the supplemental use of normative data in setting a standard. Shepard (1976), Jaeger (1978), and Conaway (1976, 1977) all favor such a procedure. Recently Jaeger (1978) advanced a standard setting method which requires judges to make judgments partially on the basis of item content. In his method, Jaeger calls for incorporation of some tryout test data to aid judges in reconsidering their initial assessments. Shepard (1976) makes the following point:

> Expert judges ought to be provided with normative data in their deliberations. Instead of relying on their experience, which may have been with unusual students or professionals, experts ought to have access to representative norms. . .of course, the norms are not automically the standards. Experts still have to decide what "ought" to be, but they can establish more reasonable expectations if they know what current performance is than if they deliberate in a vacuum.

We agree with Jaeger, Conaway, and Shepard about the usefulness of normative data when used in conjunction with a standard setting method.

6.5 Consideration of Several Promising Standard Setting Methods

Remaining methods for setting standards to be discussed in this unit assume that domain score estimates derived from criterion-referenced tests are on a continuous scale (hence, the methods fall under the heading of "Continuum Model"). For convenience, the methods under discussion are organized into three categories. The methods are presented in Figure 6.5.1. The categories are labelled "judgmental," "empirical," and "combination." In judgmental methods, data are collected from judges for setting standards, or judgments are made about the presence of variables (for example, guessing) that would effect the placement of a standard. Empirical methods require the collection of examinee response data to aid in the standard-setting process. Combination methods, not surprising, incorporate judgmental data and empirical data into the standard-setting process.

Figure 6.5.1 A classification of methods for setting standards[2]

Judgmental Methods		Combination Methods		Empirical Methods[1]	
Item Content	Guessing	Judgmental-Empirical	Educational Consequences	Data—Two Groups	Data-Criterion Measure
Nedelsky (1954)	Millman (1973)	Contrasting Groups (Zieky and Livingston, 1977)	Block (1972)	Berk (1976)	Livingston (1975)
Modified Nedelsky (Nassif, 1978)		Borderline Groups (Zieky and Livingston, 1977)			Livingston (1976)
Angoff (1971)					Huynh (1976)
Modified Angoff (ETS, 1976)					Van der Linden and Mellenbergh (1977)
Ebel (1972)					
Jaeger (1978)					

Bayesian Methods	Decision-Theoretic
Hambleton and Novick (1973)	Kriewall (1972)
Novick, Lewis, Jackson (1973)	
Schoon, Gullion Ferrara (1978)	

---

[1] Involve the use of examinee response data.

[2] From a paper by Hambleton and Eignor (1979).

6.6 Judgmental Methods

6.6.1 Item Content

In this situation, individual items are inspected, with the level of concern being how the minimally competent person would perform on the items. In other words, a judge is asked to assess how or to what degree an individual who could be described as minimally competent would perform on each item. It should be noted before describing particular procedures utilizing this criterion that while this is a good deal more objective than setting standards based on any of the methods previously discussed, a considerable degree of subjectivity still exists. Six procedures based on item content assessment will now be discussed.

i. Nedelsky Method

In Nedelsky's method, judges are asked to view each question in a test with a particular criterion in mind. The criterion for each question is, which of the response options should the minimally competent student (Nedelsky calls them "D-F students") be able to eliminate as incorrect? The minimum passing level (MPL) for that question then becomes the reciprocal of the remaining alternatives. For instance, if on a five-alternative multiple choice question, a judge feels that a minimally competent person could eliminate two of the options, then for that question, MPL = $\frac{1}{3}$. The judges proceed with each question in a like fashion, and upon completion of the judging process, sum the values for each question to obtain a standard on the total set of test items. Next, the individual judge's standards are averaged. The average is denoted $\hat{\pi}_o$.

Nedelsky felt that if one were to compute the standard deviation of individual judge's standards, this distribution would be synonomous with the (hypothesized or theoretical) distribution of the scores of the borderline students. This standard deviation, σ, could then be multiplied by a constant K, decided upon by the test users, to regulate how many (as a percent) of the borderline students pass or fail. The final formula then becomes:

$$\hat{\hat{\pi}}_o = \hat{\pi}_o + K\sigma$$

How does the K σ term work? Assuming an underlying normal distribution, if one sets K=1, then 84% of the borderline examinees will fail. If K=2, then 98% of these examinees will fail. If K=0, then 50% of the examinees on the borderline should fail. The value for K is set by (say) a committee prior to the examination.

The final result of the application of Nedelsky's method will be an absolute standard. This is because the standard is arrived at without consideration of the score distributions of any reference group. In fact, the standard is arrived at prior to using the test with the group one is concerned with testing.

The following example is included to demonstrate how the Nedelsky method can be applied in a criterion-referenced testing situation.

Example: Suppose five judges were asked to score, using the Nedelsky method, a six question criterion-referenced test made up of questions that have five response options each. Further, suppose the judges agreed that they would like 84% of the "D-F" or minimally competent students to fail (i.e., they set K=+1). The calculations below show the steps necessary to calculate a cut-off score for the test.

Judge	Test Item 1	 2	 3	 4	 5	 6	Cut-Off Score from Each Judge
A	.25	.33	.25	.25	.00	.33	1.41
B	.25	.50	.25	.50	.25	.33	2.08
C	.33	.33	.25	.33	.25	.33	1.82
D	.25	.33	.25	.33	.25	.33	1.74
E	.00	.50	.25	.33	.00	.25	1.33

Average Cut-Off Score (Across Five Judges) $= \frac{1.41 + 2.08 + 1.81 + 1.74 + 1.33}{5}$

$= 1.68$

Standard Deviation of the Cut-Off Scores $= \sqrt{\frac{(1.41-1.68)^2+(2.08-1.68)^2+\ldots+(1.33-1.68)^2}{5}}$

$= \sqrt{\frac{.380}{5}}$

$= .28$

Adjusted Cut-Off Score (84% of Borderline Student to Fail) $= 1.68 + 1 \times .28$

$= 1.96$

Therefore, approximately two test items out of six is the cut-off score on this test. From a practical standpoint, this value would seem low, but the data is created to demonstrate the process and not to model a real testing situation. Therefore, no practical significance should be attached to the answer.

ii. Modified Nedelsky

Nassif (1978), in setting standards for the competency-based teachers education and licensing systems in Georgia, utilized a modified Nedelsky procedure. A modification of the Nedelsky method was needed to handle the volume of items in the program. In the modified Nedelsky task, the entire item (rather than each distractor) is examined and classified in terms of two levels of examinee competence. The following question was asked about each item: "Should a person with minimum competence in the teaching field be able to answer this item correctly?" Possible answers were "yes," "no," and "I don't know." Agreement among judges can be studied through a simple comparison of the ratings judges give to each item. A standard may be obtained by computing the average number of "yes" responses judges give to the entire set of test items.

iii. Ebel's Method

Ebel (1972) goes about arriving at a standard in a somewhat different manner, but his procedure is also based upon the test questions rather than an "outside" distribution of scores. Judges are asked to rate items along two dimensions: Relevance and difficulty. Ebel uses four categories of relevance: Essential, important, acceptable and questionable. He uses three difficulty levels: Easy, medium and hard. These categories then form (in this case) a 3 x 4 grid. The judges are next asked to do two things:

1. Locate each of the test questions in the proper cell, based upon relevance and difficulty,
2. Assign a percentage to each cell; that percentage being the percentage of items in the cell that the minimally-qualified examinee should be able to answer.

Then the number of questions in each cell is multiplied by the appropriate percentage (agreed upon by the judges), and the sum of all the cells, when divided by the total number of questions, yields the standard.

The example that follows is modeled after an example offered by Ebel (1972).

Example: Suppose that for a 100 item test, five judges came to the following agreement on percentage of success for the minimally qualified candidate.

		Difficulty Level	
Relevance	Easy	Medium	Hard
Essential	100%*	80%	--
Important	90%	70%	--
Acceptable	90%	40%	30%
Questionable	70%	50%	20%

*The expected percentage of passing for items in the category.

BEST COPY AVAILABLE

Combining this data with the judges location of test questions in the particular cells would yield a table like the following:

Item Category	Number of Items*	Expected Success	Number X Success
ESSENTIAL			
Easy	85	100	8500
Medium	55	80	4400
IMPORTANT			
Easy	123	90	11070
Medium	103	70	7210
ACCEPTABLE			
Easy	21	90	1890
Medium	43	40	1720
Hard	50	30	1500
QUESTIONABLE			
Easy	2	70	140
Medium	8	50	400
Hard	10	20	200
TOTAL	500		37030

$$\frac{37030}{500} \doteq 74$$

---

*The number of items placed in each category by all five of the judges.

Three comments can be made about Ebel's method that should be sufficient to suggest caution when using it. One, Ebel offers no prescription for the number or type of descriptions to be used along the two dimensions. This is left to the judgment of the individuals judging the items. It is likely that a different set of descriptions applied to the same test would yield a different standard. Two, the process is based upon the decisions of judges, and while the standard could be called absolute, in that it is not referenced to score distribution, it can't be called an "objective" standard. Three, a point about Ebel's method has been offered by Meskauskas (1976):

> In Ebel's method, the judge must simulate the decision process of the examinee to obtain an accurate judgment and thus set an appropriate standard. Since the judge is more knowledgeable than the minimally-qualified individual, and since he is not forced to make a decision about each of the alternatives, it seems likely that the judge would tend to systematically over-simplify the examinees task . . . Even if this occurs only occasionally, it appears likely that, in contrast to the Nedelsky method, the Ebel method would allow the raters to ignore some of the finer discriminations that an examinee needs to make and would result in a standard that is more difficult to reach. (p. 138)

iv. Angoff's Method

When using Angoff's technique, judges are asked to assign a probability to each test item directly, thus circumventing the analysis of a grid or the analysis of response alternatives. Angoff (1971) states:

> . . .ask each judge to state the probability that the 'minimally acceptable person' would answer each item correctly. In effect, the judges would think of a number of minimally acceptable persons instead of only one such person, and would estimate the proportion of minimally acceptable persons who would answer each item correctly. The sum of these probabilities, or proportions, would then represent the minimally acceptable score. (p. 515)

v. Modified Angoff

ETS (1976) utilized a modification of Angoff's method for setting standards. Based on the rationale that the task of assigning probabilities may be overly difficult for the items to be assessed (National Teacher Exams) Educational Testing Service instead supplied a seven point scale on which certain percentages were fixed. Judges were asked to estimate the percentage of minimally knowledgeable examinees who would know the answer to each test item. The following scale was offered:

5	20	40	60	75	90	95	DNK

where "DNK" stands for "Do Not Know."

ETS has also used scales with the fixed points at somewhat different values; the scales are consistent though in that seven choice points are given. For the Insurance Licensing Exams, 60 was used as the center point, since the average percent correct on past exams centered around 60%. The other options were then spaced on either side of 60.

vi. Jaeger's Method

Jaeger (1978) recently presented a method for standard-setting on the North Carolina High School Competency Test. Jaeger's method incorporates a number of suggestions made by participants in a 1976 NCME annual meeting symposium presented in San Francisco by Stoker, Jaeger, Shepard, Conaway, and Haladyna; it is iterative, uses judges from a variety of backgrounds, and employs normative data. Further, rather than asking a question involving "minimal competence," a term which is hard to operationalize, and conceptualize, Jaeger's questions are instead:

> "Should every high school graduate be able to answer this item correctly?" "____Yes, ____No." and
> "If a student does not answer this item correctly, should he/she be denied a high school diploma?"
> "____Yes, ____No."

After a series of iterative processes involving judges from various areas of expertise, and after the presentation of some normative data, standards determined by all groups of judges of the same type are pooled, and a median is computed for each type of judge. The minimum median across all groups is selected as the standard.

Comparisons Among Judgmental Models

We are aware of two studies that compare judgmental methods of setting standards; one study was done in 1976, the other is presently underway at ETS.

In 1976, Andrew and Hecht carried out an empirical comparison of the Nedelsky and Ebel methods. In that study, judges met on two separate occasions to set standards for a 180 item, four options per item, exam to certify professional workers. On one occasion the Nedelsky method was used. On a second occasion the Ebel method was used. The percentage of test items that should be answered correctly by a minimally competent examinee was set at 69% by the Ebel method and at 46% by the Nedelsky method.

Glass (1978a) described the observed difference as a "startling finding". Our view is that since directions to the judges were different, and procedures differed, we would not expect the results from these two methods to be similar. The authors themselves report:

> It is perhaps not surprising that two procedures which involve different approaches to the evaluation of test items would result in different examination standards. Such examination standards will always be subjective to some extent and will involve different philosophical assumptions and varying conceptualizations. (p. 49)

Ebel (1972) makes a similar point:

> . . .it is clear that a variety of approaches can be used to solve the problem of defining the passing score. Unfortunately, different approaches are likely to give different results. (p. 496)

Possibly the most important result of the Andrew-Hecht study was the high level of agreement in the determination of a standard using the same method across two teams of judges. The difference was not more than 3.4% within each method. Data of this kind address a concern raised by Glass (1978a) about whether judges can make determinations of standards consistently and reliably. At least in this one study, it appears that they could. From our interactions with staff at ETS who conduct teacher workshops on setting standards, we have learned that teams of teachers working with a common method obtain results that are quite similar. And this result holds across tests in different subject matter areas and at different grade levels. We have observed the same result in our own work. Of course, certain conditions must be established if agreement among judges is to be obtained. Essentially, it is necessary that the judges share a common definition of the "minimally competent" student and fully understand the rating process they are to use.

6.6.2 Guessing and Item Sampling

In this section, some concerns initially expressed by Millman (1973) about errors due to guessing and item sampling will be discussed.

If the test items allow a student to answer questions correctly by guessing, a systematic error is introduced into student domain score estimates. There are three possible ways to rectify this situation:

1. The cut-off score can be raised to take into account the contribution expected from the guessing process.
2. A student's score can be corrected for guessing and then the adjusted score compared to the performance standard.
3. The test itself can be constructed to minimize the guessing process.

Methods one and two assume that guessing is of a pure, random nature, which is not likely to be the case for criterion-referenced tests. Thus, adjusting either the cutting score or the student's scores will probably prove to be inadequate. The test must be structured to keep guessing to a minimum, because if it occurs, it can't be adequately corrected for.

Also, if because of problems of test construction, inconvenience of administration, or a host of other problems, the test is not representative of the content of the domain, then Millman (1973) suggests that the cutting score or standard be raised (or lowered) an amount to protect against misclassification of students; i.e., false-positive and false-negative errors. Millman offers no methods for determining the extent or direction of correction for these problems. We feel that the test practitioner should exert extra effort to assure that the problem just discussed doesn't occur in the first place. Once again, there doesn't appear to be an adequate method for "correcting away" the problem.

6.7 Empirical Methods

6.7.1 Data From Two Groups

Berk (1976) presented a method for setting cut-off scores that is based on empirical data. He selects empirically the optimal cutting score for a test based upon test data from two samples of students, one of which has been instructed on the material, and the other uninstructed. Before discussing his methodology, where he offers three ways of proceeding based upon the data collected, it is worth discussing why he chose to formulate his model in the first place. He suggests that the extant approaches of a nature similar to his, namely those based on the binomial distribution and those based upon Bayesian decision-theory, suffer from a deficiency. According to Berk:

> The fundamental deficiency of all of these methods is their failure to define mastery operationally in terms of observed student performance, the objective or trait being measured, and item and test characteristics. The criterion level or cutting score is generally set subjectively on the basis of "judgment" or "experience" and the probabilities of Type 1/Type II classification errors associated with the criterion are estimated.

One of Berk's procedures considers false-positive and false-negative errors, but the difference is that the results are based upon actual data.

Berk offers three ways of approaching the problem of setting standards utilizing empirical data: (1) Classification of outcome probabilities, (2) computation of a validity coefficient, and (3) utility analysis.

i. The Basic Situation

Two criterion groups are selected for use in this procedure, one group comprised of instructed students and another of uninstructed students. The instructed group should, according to Berk, "consist of those students who have received 'effective' instruction on the objective to be assessed."

Berk suggests that these groups should be approximately equal in size and large enough to produce stable estimates of probabilities. Test items measuring one objective are then administered to both groups and the distribution of scores (putting both groups together) can be divided by a cut-off score into two categories.

Combining the classifications of students by predictor (test score) and criterion (instructed vs. non-instructed status) results in four categories that can be represented in a 2 x 2 table, with relevant marginals:

1. True Master (TM): an instructed student whose test score is above the cutting score (C).
2. False Master (FM): A Type II misclassification error where an uninstructed student's test score lies above the cutting score (C).
3. True Non-Masters (TN): An uninstructed student whose test score lies below the cutting point (C).
4. False Non-Masters (FN): Type I misclassification where an instructed student's test score lies below C.

Tabularly, this can be presented as follows. Note how the marginals are defined because they are used in the formulations to follow.

		CRITERION MEASURE	
		Instructed (I)	Uninstructed (U)
Predictor (Cutting Score)	Predicted Masters PM=TM+FM	(TM)	Type II (FM)
	Predicted Non-Masters PN=FN+TN	Type I (FN)	(TN)
		Masters M=TM+FN	Non-Masters N=FM+TN

BEST COPY AVAILABLE

ii. Classification of Outcome Probabilities

In this procedure, identification of the optimal cutting score involves an analysis of the two-way classification of outcome probabilities shown above. This can be done algebraically by following the steps listed below, or graphically, as illustrated in a subsequent section. The steps to follow are:

1. Set up a two-way classification of the frequency distribution for each possible cutting score.
2. Compute the probabilities of the 4 outcomes (for each cutting score) by expressing the cell frequencies as proportions of the total sample. For instance:

   Prob (TM) = TM/(M+N)

   Prob (FM) = FM/(M+N)

   Prob (TN) = TN/(M+N)

   Prob (FN) = FN/(M+N)
3. For each cutting score, add the probability of correct decisions: Prob (TM) + Prob (TN), and the probability of incorrect decisions: Prob (FN) + Prob (FM).
4. The optimal cutting score is the score that maximizes Prob (TM) + Prob (TN) and minimizes Prob (FN) + Prob (FM). It is sufficient to observe the score that maximizes Prob (TM) + Prob (TN) because [Prob (FN) + Prob (FM)] = 1 - [Prob (TM) + Prob (TN)]. That is, the score that maximizes the probability of correct decisions automatically minimizes probability of incorrect decisions.

iii. Graphical Solution

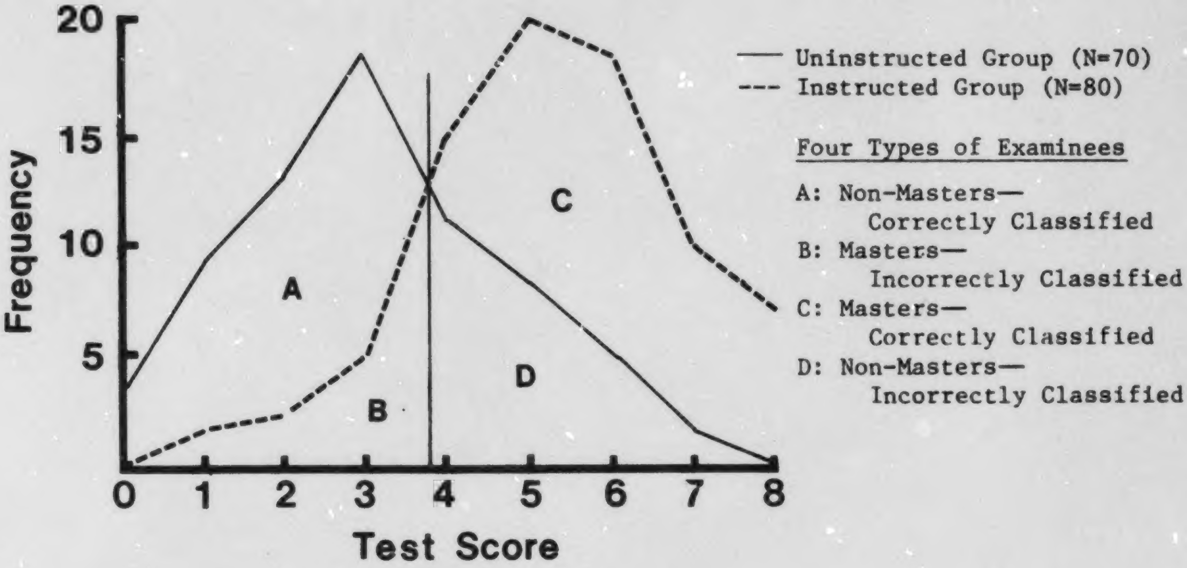
Berk (1976) also mentions that the optimal cutting point for a criterion-referenced test can be located by observing the frequency distributions for the instructed and uninstructed groups. According to Berk:

> The instructed and uninstructed group score distributions are the primary determinants of the extent to which a test can accurately classify students as true masters and true non-masters of an objective. The degree of accuracy is, for the most part, a function of the amount of overlap between the distribution.

If the test distributions overlap, no decisions can be made. The ideal situation would be one in which the two distributions have no overlap at all. A typical situation we should hope for is for the instructed group distribution to have a negative skew, the uninstructed group to have a positive skew, and for there to be a minimum of overlap. The point at which the distributions intersect is then the optimal cut-off score.

In Figure 6.7.1, the distributions of test scores for two groups of examinees (one instructed group and one uninstructed group) are shown.

Figure 6.7.1 Frequency polygons of criterion-referenced test scores for two groups - an instructed group and an uninstructed group on the content measured by the test.



Frequency Distribution of Test Scores

Test Score	U I Group	I Group
8	0	7
7	2	10
6	5	18
5	8	20
4	11	15
3	18	5
2	13	3
1	9	2
0	4	0

iv. Validity Coefficient

In this procedure, a validity coefficient is computed for each possible cutting score. The cutting score yielding the highest validity coefficient also yields the highest probability of correct decisions. To utilize the procedure, the following steps should be followed:

1. From the two-way classification introduced earlier, compute the base rate (BR) and the selection ratio (SR). They are given by:

   BR = Prob (FN) + Prob (TM)

   SR = Prob (TM) + Prob (FM)

2. Calculate the phi coefficient $\emptyset_{vc}$ using the following formula:

$$\emptyset_{vc} = \frac{\text{Prob (TM)} - \text{BR (SR)}}{\sqrt{\text{BR (1-BR) SR(1-SR)}}}$$

3. The cutting score yielding the highest $\emptyset_{vc}$ is the optimal cutting score.

The formula for the phi coefficient, $\emptyset_{vc}$, given above is suitable for a 2 x 2 table of cell probabilities. More generally, the phi coefficient is the Pearson product moment correlation between two dichotomous variables, and could be arrived at as follows:

1. Each student with a test score above the cutting score in question is assigned a 1, below a 0.
2. Each student in the instructed group is assigned a 1, in the uninstructed group, a 0.
3. $\emptyset_{vc}$ would then be the correlation coefficient computed in the usual way.

V. Utility Analysis

In this section, costs or losses are assigned to the misclassification of students as false masters or false non-masters. The procedures here are closely tied to the decision-theoretic procedures discussed in a later section. The procedure is presented at this point because it can be related to the two Berk procedures just discussed.

First of all, Berk notes the following fact:

> When the outcome probabilities or validity coefficient approach is used to select the optimal cutting score, it is assumed that the 2 types of errors are equally serious. If, however, this assumption is not realistic in terms of the losses which may result from a particular decision, the error probabilities need to be weighted to reflect the magnitude of the losses associated with the decision.

Berk notes that determination of the relative size of each loss is judgmental, and must be guided by the consequences of the decision considered. He mentions considering the following factors: Student motivation, teacher time, availability of instructional materials, content, and others. Berk suggests the following, which we have capsulized into a series of steps:

1. Estimate the expected disutility of a decision strategy (ζ) by

   ζ_k = Prob (FN) $[D_1]$ + Prob (FM) $[D_2]$

   where D_1 and $D_2 < 0$

   and k = the single decision in question

   D_1 and D_2 = respective disutility values

2. Estimate the expected utility of a decision strategy (ν) by

   ν_k = Prob (TM) $[U_1]$ + Prob (TN) $[U_2]$

   where U_1 and $U_2 > 0$

   and k = the single decision in question (same as for disutility)

   U_1 and U_2 = respective utility values

3. Form a composite measure of test usefulness by combining the estimates of utility and disutility across all decisions

$$\gamma = \sum_{k=1}^{n} (\nu_k + \zeta_k)$$

γ = index of expected maximal utility.

4. Choose the cutting score with the highest γ index (it maximizes the usefulness of the test for decisions with a specific set of utilities and disutilities).

vi. Suggestions

The procedures developed by Berk (1976) hold considerable promise for use in setting criterion-referenced test score standards. The ideas in his procedures are now new; there are other procedures that are concerned with the maximization of correct decisions and the minimization of false-positive and false-negative errors. The attractive feature is the ease with which Berk's methods can be understood and applied. The major potential drawback is in the assignment of examinees to criterion groups. If many examinees in the "instructed group" do not possess the assumed knowledge and skills measured by the criterion-referenced test (or if many examinees in the "uninstructed group" do), Berk's methods will produce inaccurate results.

6.7.2 Decision-Theoretic Procedures

Berk (1976) looked at the minimization of false-positive and false-negative decisions through the use of actual test data. He selects as optimal the cutting score that minimizes false-positive and false-negative errors. Another way to look at false-positive and false-negative errors is to assume an underlying distributional form for your data and then observe the consequences of setting values, such as cutting points, based upon the distributional model. The logic is the same here in terms of minimization of errors, except that by assuming a distributional form, actual data does not have to be collected. Situations can be simulated or developed, based upon the model.

Meskauskas (1976) has related and compared these procedures to those based upon analyses of the content of the test. In reference to these models, of which we will describe one:

> . . .the models to follow deal with approaches that start by assuming a standard of performance and then evaluating the classification errors resulting from its use. If the error rate is inappropriate, the decision maker adjusts the standard a bit and tries his equation again.

Before discussing one of the procedures in greater detail, the Kriewall binomial-based model, the procedures discussed here should be related to criterion-referenced testing procedures involving the determination of test length. Many of the test length determination procedures (Millman, 1973; Novick & Lewis, 1974) make underlying distributional assumptions and proceed in the fashion discussed above by Meskauskas. The focus of concern, however, is test length determination, and not the setting of a cutting score. In fact, Millman's (1973) procedure is based upon exactly the same underlying distribution, the binomial, as is Kriewall's model to be discussed. It should be pointed out that the procedures are exactly the same, the data is just represented differently because of the level of concern, either cutting score or test length.

i. Kriewall's Model

Kriewall's (1972) model focuses on categorization of learners into several categories: Non-master, master, and an in-between state where the student has developed some skills, but not enough to be considered a master.

Kriewall assumes the function of measurement, using the test, is to classify students into one of two categories, master or non-master. Of course, the test, as a sample of the domain of tasks, is going to misclassify some individuals as false-positives (masters based on the test, but non-masters in reality) and false-negatives (non-masters on the test, but masters in reality). By assuming a particular distribution, these errors may be studied.

Kriewall's probability model, used to develop the likelihood of classification errors, is based upon the binomial distribution. He assumes:

1. The test represents a randomly selected set of dichotomously scored (0-1) items from the domain.
2. The likelihood of correct response for a given individual is a fixed quantity for all items measuring a given objective.
3. Responses to questions by an individual are independent. That is, the outcome of one trial (taking one question) is independent of the outcome of any other trial.
4. Any distribution of difficulty of questions (for an individual) within a test is assumed to be a function of randomly occurring erroneous responses (Meskauskas, 1976).

With these assumptions, Kriewall views a student's test performance as "a sequence of independent Bernoulli trials, each having the same probability of success." A sequence of Bernoulli trials follows a binomial distribution, which has a probability function which relates the probability of occurrence of an event (a particular test score) to the number of questions in the test by:

$$f(x) = \binom{n}{x} p^x q^{n-x} \quad ,$$

where

x = a test score

n = total number of test items

p = examinee domain score

$q = 1-p$

and

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} \quad .$$

Kriewall sets some boundary values and a cutting score, and then looks at the probability of misclassification errors. Using the notation of Meskauskas (1976), set:

β_1 = the lower bound of the mastery range (as a proportion of errors)

β_2 = the upper bound of the non-mastery range

C = the cutting score; the maximal number of allowable errors for masters. Kriewall recommends $C = \frac{\beta_1 + \beta_2}{2}$

Given values for the above three variables, Kriewall uses the (assumed) binomial distribution to determine the probabilities. If α is the probability of a false positive result (a non-master who scores in the mastery category) and β is the probability of a false negative result (a master who scores in the non-mastery category), then α and β are given by:

$$\alpha = \sum_{w=c}^{n} \binom{n}{w} z_1^{n-w} (1 - z_1)^w$$

$$\beta = \sum_{w=o}^{c-1} \binom{n}{w} z_2^{n-w} (1 - z_2)^w$$

where w = observed number of errors (and w = n-x) for an individual.

According to Meskauskas (1976) the formula for α is:

> . . .equivalent to obtaining the probability that, given a large number of equivalent trials, a person whose true score is equal to the lowest score in the mastery range will fall in the non-mastery range.

By setting z_1 and z_2 at various values, and determining $C = \frac{z_1 + z_2}{2}$, the probabilities of false positive and false negative errors can be studied. The optimal value for C (and thus z_1 and z_2) would then be the value that minimized α and β. The results are dependent, however, on n and w.

ii. Suggestions

While Kriewall has offered a method of studying classification errors that does not depend upon actual data, we prefer the method of Berk, due to its simplicity. Kriewall's model seems to us to fit in much better with the procedures on test length determination. For instance, suppose you have specified minimal values for α and β, and have determined C, the cutting point. Then the formulas above for α and β can be solved for n, the total number of questions needed. (It would be much easier if one isolated n on the left hand side). This is exactly what is done when using the binomial model to solve the test length problem.

In sum, we prefer the Berk method for observing probabilities of misclassification errors both because of its simplicity and because of the lack of restricting underlying distributional assumptions. Kriewall's method does, however, offer a viable alternative for setting a cut-off score when actual test data cannot be collected.

6.7.3 Empirical Models Depending Upon a Criterion Measure

The models to be discussed in this section bear great resemblance to both Berk's and Kriewall's methods just discussed. They have been separated from those two methods because these methods are built upon the existence of an outside criterion measure, performance measure, or true ability distribution. The test itself, and the possible cut-off scores, are observed in relationship with this outside measure. The optimal cut-off is then chosen in reference to the criterion measure. For instance, Livingston's (1975) utility-based approach leads to the selection of a cut-off score that optimizes a particular utility function. The procedure of Vander Linden and Mellenburgh (1976), in contrast, leads to the selection of a cut-off score that minimizes expected loss.

In reference to the setting of performance standards based upon benefit (and cost) Millman (1973) has suggested that psychological and financial costs be considered:

> All things being equal, a low passing score should be used when the psychological and financial costs associated with a remedial instructional program are relatively high. That is, there should be fewer failings when the costs of failing are high. These "costs" might include lower motivation and boredom,

BEST COPY AVAILABLE

> damage to self-concept, and dollar and time expenses of conducting a remedial instructional program. A higher passing score can be tolerated when the above costs are not too great or when the negative effects of moving a student too rapidly through a curriculum (i.e., confusion, inefficient learning and so forth) are seen as very important to avoid.

In sum, to utilize these procedures, a suitable outside criterion measure must exist. Success and failure (or probability of success and failure) is then defined on the criterion variable and the cut-off chosen as the score on the test that maximizes (or minimizes) some function of the criterion variable. The existence of such a criterion variable has implications for the utilization of these methods for setting cut-off scores on minimum competency tests.

i. Livingston's Utility-based Approach

Livingston (1975) suggests the use of a set of linear or semi-linear utility functions in viewing the effects of decision-making accuracy based upon a particular performance standard or cut-off score. That is, the functions relating benefit (and cost) of a decision are related linearly to the cutting score in question.

Livingston's procedure is like Berk's procedure for utility analysis discussed in 6.7.1 except that Livingston develops his procedure based upon any suitable criterion measure (not just instructed versus uninstructed), and also specifies the relationship between utility (benefit or loss) and cutting scores as linear. The relationship does not have to be linear; however, using such a relationship simplifies matters somewhat. In such a situation the cost (of a bad decision) is proportional to the size of the errors made and the benefit (of a good decision) is proportional to the size of the errors avoided.

ii. Van der Linden and Mellenburgh's Approach

The developers of this procedure have prescribed a method for setting cutting scores that is related both to Berk's procedure and Livingston's. We will describe the procedure briefly and in the process relate it to Berk's work. A test score is used to classify examinees into two categories: Accepted (scores above the cutting score) and rejected (scores below). Also, a latent ability variable is specified in advance and used to dichotomize the student population: Students above a particular point on the latent variable are considered "suitable" and below "not suitable." The situation may be represented as follows.

		Latent Variable	
		Not suitable $\gamma < d$	Suitable $\gamma \geq d$
Decision	Accepted $X \geq C$	"False +" $\ell_{01}(\gamma)$	$\ell_{11}(\gamma)$
	Rejected $X < C$	$\ell_{00}(\gamma)$	"False -" $\ell_{10}(\gamma)$

where C = cutting score on the criterion-referenced test

d = cutting score on the latent variable ($0 \leq d \leq 1$),

and where ℓ_{ij} ($i,j = 0,1$) is a function of γ and related in the general loss function:

$$L = \begin{cases} \ell_{00}(\gamma) & \text{for } \gamma < d,\ X < C \\ \ell_{10}(\gamma) & \text{for } \gamma \geq d,\ X < C \\ \ell_{01}(\gamma) & \text{for } \gamma < d,\ X \geq C \\ \ell_{11}(\gamma) & \text{for } \gamma > d,\ X \geq C \end{cases}$$

BEST COPY AVAILABLE

The authors then specify risk (the quantity to be minimized) as the expected loss, and the cutting score that is optimal is the value of C that minimizes the risk function (expected value of loss). They simplify matters (as does Livingston) by specifying their loss function as linear.

In sum, while Van der Linden and Mellenburgh have provided a method for setting a cut-off score on the test, they have offered little to help in setting the cut-off on the latent variable. In a sense then, they have only transferred the problem of setting a standard to a different measure!

iii. Livingston's Use of Stochastic Approximation Techniques

Livingston (1976) has developed procedures for setting cut-off scores based upon stochastic approximation procedures. According to Livingston, the problem involving cut-off scores can be phrased as follows to fit stochastic procedures: "In general, the problem is to determine what level of input (written test score) is necessary to produce a given response (performance), when measurements of the response are difficult or expensive." The procedure, according to Livingston, is as follows:

1. Select a person; record his/her test score and measure his/her performance.
2. If the person succeeds on the performance measure (if his/her performance is above the minimum acceptable), choose next a person with a somewhat lower test score. If the person fails on the performance measure, choose a person with a higher written test score.
3. Repeat step 2, choosing the third person on the basis of the second person's measured performance.

Livingston offers two different procedures for choosing step size, the up-and-down and the Robbins-Monro Procedure, and a number of procedures for estimating minimum passing scoresconsonant with each.

This procedure, like those discussed earlier in this section, depends upon the existence of a cut-score established on another variable, this time the performance measure, in order to establish the passing score on the test. This then limits greatly the applicability of the method. Livingston (personal communication, 1978) has suggested that judgmental data on performance can be used, rather than actual performance data, with the procedure, but this has yet to be documented in any fashion. When documented, the possibilities for use of the procedures will be greatly expanded.

iv. Huynh's Procedures

Huynh (1976) has advanced procedures for setting cut-off scores that are predicated on the existence of a "referral task." This referral task can be envisioned as an external criterion to which competency can be related. For instance, Huynh (1976) states that "Mastery in one unit of instruction may not be reasonably declared if it cannot be assumed that the masters would have better chances of success in the next unit of instruction." The next unit in this case would be the referral task.

These procedures once again depend upon an outside criterion variable to permit the estimation of a cut-score. In

BEST COPY AVAILABLE

this case, the user of the method is asked to establish the probability of success of individuals on the referral task. Because of the necessity of a criterion variable for operation, these procedures suffer in generalizability. They are, for instance, apparently not useful for minimum competency testing situations where a criterion variable, and associated probability of success, are next to impossible to establish.

6.7.4 Educational Consequences

In this situation, one is concerned with looking at the effect setting a standard of proficiency has on future learning or other related cognitive or affective success criteria. According to Millman (1973), the question here "What passing score maximizes educational benefits?".

This approach can be visualized from an experimental design point of view. A subject matter domain is taught to a class of students who are then tested on the material. These students are assigned (randomly) to groups with the groups differing on the performance level required for passing the test. The students are then assessed on some valued outcome measure and the level of performance on the criterion-referenced test for which the valued outcome is maximal (it could be a combination of valued outcomes) becomes the performance standard or criterion score.

Thus, to use this method, much more data needs to be collected than for the item content procedures. An experiment must be conducted, and then a cut-off score is selected based upon the results of the experiment.

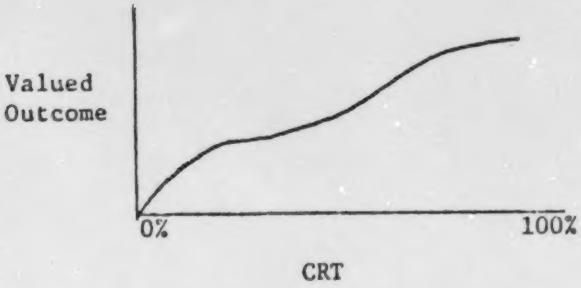
Because of the difficulties involved in designing and carrying out experiments in school settings, the method is unlikely to find much use.

1. Block's Study

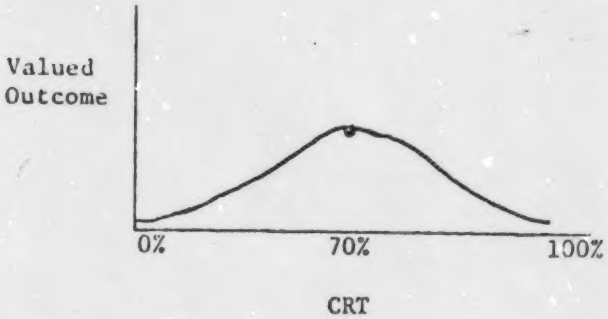
Block's study (1972) involves students learning a subject segment on matrix algebra using a Mastery Learning paradigm. Such a paradigm dictates that students who don't perform adequately on the posttest be recycled through remedial activities until they demonstrate mastery (re: attain a score above the cutting score). Block established four groups of students, where each group was tested using one of the following four performance standards: 65, 75, 85, and 95% of the material in a unit must be mastered before proceeding on the next unit. He then examined the effects of varying the performance standard on six criteria that were used as the variables to be maximized. Viewing these criteria as either cognitive or affective, Block observed that the 95% performance level maximized student performance on the cognitive criteria, while the 85% performance level seemed to maximize the affective criteria.

Some comments on Block's study are in line. One, the results lack generalizability. The 95% and 85% levels, which maximize the cognitive and affective measures respectively, are likely to change with the subject matter. Two, as pointed out by Glass (1978a), the method of maximizing a valued outcome assumes that there is a distinct point or criterion score on the CRT that maximizes the outcome. What if the curve relating performance on the CRT is monotonically increasing, so that 100% performance on the CRT maximizes the valued outcome? In fact, it is more likely to be the case that the graph is monotonically increasing than the case where the graph increases and decreases. For example:

1. Monotonically increasing graph (Problem situation)



2. Ideal situation



(Reproduced from Glass, 1978a, permission for reproduction pending.)

Thus, it can be seen that unless the graph increases and then decreases, a 100% performance standard will be optimal. This standard is of limited use because it is not realistic to expect all students to attain that level.

Third, Block discusses that if there are multiple criteria to be maximized as valued outcomes, then some model for combining criteria with relevant weights needs to be developed. He does not offer any procedures for doing so however, and he looks at the effects of the performance standards on each of the 6 criteria separately. It should be noted that multiple criteria is a way around the problem discussed above (Glass, 1978a). For instance, if one of the outcomes has a monotonically increasing relationship with the test scores and the other a monotonically decreasing relationship, then the composite should have a peak value at a point other than 0% or 100%. While this would seem to solve the problem, another problem is only further exacerbated; what weights should be assigned to the valued outcomes to form the composite? These procedures have not yet been developed, and further, they are likely to be situation specific.

BEST COPY AVAILABLE

6.8 Combination Methods

6.8.1 Judgmental-Empirical

Zieky and Livingston (1977), and more recently, Popham (1978), have suggested two procedures that are based upon a combination of judgmental and empirical data. In addition, both Zieky and Livingston and Popham have included an in-depth discussion of how to implement the procedures, something that has been lacking with many other procedures. The two procedures presented by Zieky and Livingston, the Borderline-Group and Contrasting-Groups methods, are procedurally similar. They differ in the sample of students on which performance data is collected. Further, while judgments are required, the judgments necessary are on students; not on items, as are many of the other judgmental methods (Nedelsky, Angoff, Ebel, etc.). Zieky and Livingston make the case that judging individuals is likely to be a more familiar task than judging items. Teachers are the logical choice as judges, and for them, the assessment of individuals is commonplace.

i. Borderline—Group Method

This method requires that judges first define what they would envision as minimally acceptable performance on the content area being assessed. The judges are then asked to submit a list of students (about 100 students) whose performances are so close to the borderline between acceptable and unacceptable that they can't be classified into either group. The test is thus administered to this group, and the median test score for the group is taken as the standard.

ii. Contrasting-Group Method

Once judges have defined minimally acceptable performance for the subject area being assessed, the judges are asked to identify those

BEST COPY AVAILABLE

students they are sure are either definite masters or non-masters of the skills measured by the test. Zieky and Livingston suggest 100 students in the smaller group in order to assure stable results. The test score distributions for the two groups are then plotted and the point of intersection is taken as the initial standard. This is exactly the same as the graphical procedure suggested by Berk, and presented in section 6.7.1. Zieky and Livingston then suggest adjusting the standard up or down to reduce "false masters" (students identified as masters by the test, but who have not adequately mastered the objectives) or "false non-masters" (students identified as non-masters by the test, but who have adequately mastered the objectives). The direction to move the cut-off score depends on the relative seriousness of the two types of errors.

iii. Suggestions

These methods, particularly the Contrasting-Groups Method, are very similar to the procedure suggested by Berk. Instead of actually forming instructed and uninstructed groups, however, as suggested by Berk, the Contrasting-Groups Method asks judges to form the groups. This judgmental procedure would seem more advantageous when the content being assessed has had a long instructional period (minimum competency testing is an example), or when there would be problems justifying the existence of an uninstructed group. Berk's method would be more useful for tests based on short instructional segments, most likely administered at the classroom level.

A comparison of the judgments involved in the two procedures indicates that the Contrasting-Groups Method would be the easier method to justify using. It is a more reasonable task for teachers to identify "sure" masters and non-masters than it is for them to identify borderline students in the subject area being assessed. In sum, the Contrasting-Groups Method appears to us to be a most reasonable way of setting standards.

6.8.2 Bayesian Procedures

Novick and Lewis (1974) were the first to suggest that Bayesian procedures are useful for setting standards. Schoon, Gullion, and Ferrara (1978) have more recently discussed Bayesian procedures for setting standards. According to Schoon et al., Bayesian procedures allow the incorporation of:

1. A loss ratio, reflecting the severity of false-positive and false-negative decision errors,
2. prior information on the distribution of domain scores in the population of interest,
3. current information on an examinee's domain score, and
4. the degree of certainty that an examinee's domain score exceeds the cut-off score.

Of course, a cut-off score must first be set in order for the four factors to be incorporated. Thus, Bayesian procedures offer a way of augmenting the establishment of a cut-off score rather than a method for setting the cut-off score itself.

In sum, Bayesian procedures present a method for augmenting the setting of a cut-off score by utilizing available prior and collatoral information. The procedure also provides a posterior statement of degree of certainty about candidate's performance. Bayesian procedures do not, however, offer a method for setting a cut-score in the first place. Bayesian procedures have been included in this review because they do offer a method for combining judgmental and empirical data to arrive at a revised standard.

6.9 Some Procedural Steps in Standard Setting

In earlier sections of this unit, issues and many methods for standard-setting were discussed. In this section, procedures will be outlined for setting standards on criterion-referenced tests used for three different purposes. The purposes considered are:

1. Classroom testing
2. Basic skills testing for yearly promotion and high school graduation
3. Professional licensing and certification testing.

Classroom testing is emphasized since classroom teachers have fewer technical resources available to them than do the larger testing programs. Our ultimate objective is to provide a comprehensive set of practical guidelines for practitioners. At this time the guidelines are far from comprehensive; much research is needed to supply information necessary to construct thorough guidelines. We have suggested in places some of the questions that need to be answered.

Certain things are assumed: first, that in each case a set of objectives or competencies has been agreed upon, and that they are described via the use of domain specifications or some other equally appropriate method. Second, it is assumed that no fixed selection ratio exists (e.g., one might be fixed in effect by having resources to provide only a certain number of students with remedial work) since if it does there is no reason to set standards. Finally, we do not discuss the important and interesting political issues of who participates in and who controls the standard-setting process; we take as given that some such process exists and only address the issue of participation from the perspective of practicality.

6.9.1 Preliminary Considerations

Before any standard setting is undertaken for any purpose, an analysis of the decision-making context and of the resources available for the project should be done. The results of this analysis will determine how extensive and sophisticated the standard-setting procedure should be. Analysis of the decision-making context involves judging the importance of the decisions that are to be made using the test, the probable consequences of those decisions, and the costs of errors. Others have discussed using these same considerations in adjusting the final standard, but they may also be helpful in choosing a standard-setting method. Formal procedures for using this information are probably not necessary; a discussion of the issues by those directing the project should suffice. Some issues to consider would include (1) the number of people directly and indirectly affected by the decisions to be based on the test; (2) possible educational, psychological, financial, social and other consequences of the decisions; and (3) the duration of the consequences.

The next step should be a consideration of the resources available for the standard setting. Resources include money, materials, clock time, personnel time and expertise. How much of the total amount of available resources will be dedicated to the standard setting will depend upon the results of the prior discussion of decision context. The final decision as to the resources to be invested will determine how large and technically sophisticated the standard-setting enterprise may be.

A great deal of information needs to be collected on the actual expenditures of various resources that have been required to carry out standard setting by different methods in different contexts. Actual time and money data would be invaluable to practitioners in choosing a method for their own situation. In the following discussion procedural steps in increasing order of expense and complexity will be offered but real data on these factors is lacking and is a pressing need.

6.9.2 Classroom Testing

The classroom teacher is most likely to use criterion-referenced tests for diagnostic purposes, that is for determining whether a student has mastered an area or needs further work in it. This would seem to be the most common situation calling for the setting of standards. Here the teacher must decide what level of test performance constitutes "mastery." In the same testing context the teacher may set additional performance standards, above and/or below the minimal level, for the awarding of grades on the material.

Typically the classroom teacher works alone, or at most with one or more other teachers of the same grade. It is also quite often the case that a classroom exam is used only once. In these situations methods based only on judgment of test content may be the only ones practicable. The methods developed by Ebel, Nedelsky and Angoff would be appropriate here, and the details of each of them have been discussed in an earlier section, so we will not re-iterate procedural steps here.

When available resources permit involving more people in the standard setting, parents and other community members might be enlisted, or a group of teachers of one grade from an entire school district might collaborate in setting standards. Again, if resources permit, data on group performance on individual items may be tabulated and considered in setting the standards on subsequent tests, or if tests are retained from year to year, the performance data from the previous year might be used. Of course, this can also be done by teachers working alone. The following is a list of steps, some of which could be omitted if resources were limited, for involving parents of students in a particular class in setting standards for classroom tests over units of instruction. The method borrows heavily from Jaeger (1978). (It is assumed that the objectives have been identified and the teacher (or teachers) has prepared domain specifications):

1. At the beginning of the school year, a letter is sent to parents explaining the project and inviting them to a meeting where more information will be given.
2. At the meeting parents are given copies of domain specifications for the first test, along with example items. They are asked to indicate for each objective a percentage of items, which answered correctly would demonstrate the student had mastered the material adequately. At this meeting they should be encouraged to discuss the task and ask any questions they might have about it.

Instructions accompanying the standard-setting task should indicate to the parents how their judgment will be employed (for example, averaged with the percentages indicated by every other parent, and the resulting standard applied to every child in that class or grade). We have suggested for reasons of test security that the parents base their judgments on domain specifications rather than on actual test items; if test forms from previous years are available and thought to be parallel to the new exam, it may be easier for parents to make their judgments as a percentage correct of items on the parallel test.

3. The teacher constructs the criterion-referenced test from the domain specifications before looking at the parents' standards.
4. Class performance data is tabulated after the test is administered.

5. Parent judgment for the second test (or set of tests) is solicited by mail. The mailing packet includes: domain instructions (duplicating those given at the earlier meeting), and performance data from the first test (number of students achieving each set standard).

Instructions would also stress that judgments were to be based primarily on domain specifications and only secondarily on performance data.

6. Step 5 is repeated during the year whenever a competency-type test is to be given.

Alternatively, this procedure might be reserved for those instructional units judged to cover basic, required objectives for that grade; parents' instructions would then identify the tested materials as such.

7. The teacher keeps files for each test, including the domain specifications, parent judgment forms, actual exam and performance data.

8. Periodic meetings can be held to review the instructions and to discuss the procedure and its results.

Such discussions may lead to parents questioning the performance of students, and is likely to provoke query into both the teacher's methods and his/her subject matter. Teachers should be prepared for this; it may lead to parents wanting greater involvement in determining other aspects of their children's schooling, a desire one hopes can be creatively and constructively used.

Other variants on this procedure can include appointing a small committee of parents, possibly working with several teachers, instead of an open parents group. A parent-objective (matrix) sampling strategy could be employed to reduce the number of judgments required of each parent.

Another procedure for setting standards with criterion-referenced tests in instructional settings was offered by Hambleton (1978). According to Hambleton, "[His] is not a 'validated list' of guidelines. It is a list of practical guidelines I have evolved over the years through my work with numerous school districts." His eleven step list of guidelines is as follows:

1. The determination of cut-off scores should be done by several groups working together. These groups include teachers, parents, curriculum specialists, school administrators, and (if the tests are at the high school level) students. The number from each group will depend upon the importance of the tests under consideration and the number of domain specifications. At a minimum, I like to have enough individuals to form at least two teams of reviewers. This way I can compare their results on at least a few domain specifications to determine the consistency of judgments in the two groups. When sufficient time is available I prefer to obtain two independent judgments of each cut-off score.

2. I usually introduce either the Ebel method or the Nedelsky method. Following training on one of the methods, I have the groups work through several practice examples. Differences between groups are discussed and problems are clarified.

3. The domain specifications (or usually, but less appropriate, the objectives) are introduced and discussed with the judges.

4. I try to set up a schedule so that roughly equal amounts of time are allotted to a consideration of each domain specification. If some domain specifications are more complex or important I usually assign them more time.

5. I make sure that the judges are aware of how the tests will be used and with what groups of students.

6. If there exist any relationships among the domain specifications (or objectives) the information is noted. For example, if a particular objective is a prerequisite to several others it may be desirable to set a higher cut-off score than might otherwise be set.

7. Whenever possible I try to have two or more groups determine the cut-off scores. Consistency of their ratings can be studied, and when necessary, differences can be studied, and a consensus decision reached.

8. If some past test performance data are available, it can be used to make some modifications to the cut-off scores. On some occasions, instead of modifying cut-off scores, decisions can be made to spend more time in instruction to try and improve test performance. If past group performance on an objective is substantially better than the cut-off score, less time may be allocated to teaching the particular objective.

9. As test data become available, percentage of "masters" and "non-masters" on each objective should be studied. If performance on some objectives appears to be "out of line," an explanation can be sought by a consideration of the test items (perhaps the test items are invalid), the level of the cut-off score, variation in test performance across classes, a consideration of the amount of instructional time allotted to the objective and so on.

10. Whenever possible I try to compare the mastery status of uninstructed and instructed groups of examinees. Instructed groups ought to include mainly "master" students. The uninstructed groups should include mainly the "non-masters." If many students are being misclassified, a more valid cut-off score can sometimes be obtained by moving it (for example, see Berk, 1976).

11. It is necessary to re-review cut-off scores occasionally. Curriculum priorities change and so do instructional methods. These shifts should be reflected in the cut-off scores that are used.

There are many important questions needing to be researched. These techniques have apparently been used very little (there is certainly much more literature on how to set standards than on what happens when one does); we need to know the effects of involving different groups of people in the standard-setting (especially parents as opposed to others), of the number of people involved, the information and instructions provided and the frequency of standard setting. How do these factors effect the levels set, the public acceptability of the chosen standard, and are the procedures cost-effective?

6.9.3 Basic Skills Testing for Annual Promotion and High School Graduation

These are clearly areas where greater importance is attached to the consequences of testing and, hence, more resources will be allocated than for classroom testing. The discussion is limited here to testing of "minimal" competencies, not intending that the procedures be applied to the total curriculum. Further, we are not discussing the "life skill" or "survival" competencies; in setting standards for these skills it is necessary to consider performance on criterion measures of life success. We feel that this undertaking is beyond the capabilities of educational and measurement practice. It will be difficult enough to decide upon and assess "minimal" skills. For these skills, since no external criterion measures can be said to exist, the appropriate performance data to consider in standard setting are scores on the actual tests (or items). We agree with those (e.g., Jaeger, 1978; Linn, 1978; Shepard, 1976) who hold that performance data should be considered along with test content to inform the setting of standards. While from an idealistic point of view it would be desirable to set standards with reference only to the content of a domain, in reality the degree of skill in test construction required for the pure-content approach is probably beyond human attainment. In order to avoid unpleasant shocks it would seem good practice to examine test performance data; the other benefit of so doing is that feedback is received on our content-based judgments and may thus refine our skills.

Jaeger (1978) has provided an excellent guide to implementing a procedure involving representative groups affected by standards set for high school graduation. The method was discussed earlier, but a brief review at this point seems useful. In general terms, it is an iterative procedure for soliciting item-by-item judgments from groups of judges. Information fed back to the judges at each iteration includes (a) group performance on each test item in a pilot administration, (b) the percentage of students who would have passed given several different standards, and (c) a distribution of the standards suggested by the judges in the group. The median passing score for each type of judge is computed, and the lowest of the medians taken as the standard.

The principal attraction of plans such as Jaeger's and the one outlined in Section 6.9.2, which is based on Jaeger's, is their political viability. By involving a broad cross-section of constituents in the setting of the standard, one increases the acceptability of that standard. However, no actual control or very significant influence over the educational process is transferred to the constituency; the objectives and the test, after all, are presented to them as givens, and their contribution in setting the standard is really quite limited. Moreover, the consensus method, while probably not harmful, may not produce results that make any pedagogical sense. Where obtaining popular support is not a critical problem, educators may prefer to rely upon the judgments of subject-matter and measurement "experts" to set standards. This may produce a more coherent, if less universally-accepted, result. Such a procedure could be implemented as follows (the steps would be executed for each subject matter area by content experts working with measurement experts):

1. Categorize the educational objectives or competencies as being of the knowledge/information type or of the rule-learning type (this distinction corresponds to Meskauska's (1976) continuum vs. state mastery models).

In the first case it makes sense to speak of a domain score, and to sample randomly from the domain to estimate that score. In the second, since learning is presumed to be all-or-none, sampling considerations are not relevant, but construction of a few test items that accurately reflect the ability is critically important. Objectives domains of the first type reflect Ebel's (1978) notion of the purpose of competency certification tests as being efficient and accurate indicators of the level of achievement in a broad domain, rather than lists of specific competencies attained.

2. For objectives or competencies of the first type, construct tests with the aid of domain specifications, items matched to the domain specifications, and a suitable item sampling plan.

3. Ebel's standard-setting method (or one of the other content-focused methods) may then be used to set the standard for these parts of the test. To use Ebel's method the items from all of the knowledge/information (or continuum) domains would be considered together. (Table 6.9.3 provides a comparison of six possible methods.)

4. Pooling the judgments of all the experts may present a problem. Simply averaging the ratings given to each item (on relevance and difficulty) and/or the standards assigned to each category, will probably not give a very meaningful result. Ideally, the experts will go through a series of iterations in which they compare their independent judgments (first of the item categorization and next of the standards they assigned to each category), note discrepancies, discuss the rationale for each judgment, possibly decide upon revisions in the test (this will direct the procedure back to Step 2, to ensure that any revisions do not distort the test's domain representativeness), and/or persuade each other to change their judgments. Unanimity might be required in order to proceed from this step.

5. For those objectives or competencies classified as being of the "State" variety, smaller sets of items are required since the domains are more homogeneous, but item construction must be, if anything, more painstaking. Ideally, experimental evidence would be garnered to show that item performance truly reflected the target construct.

6. Standards on these State-type objectives can be adjusted back from 100% using Emrick's (1971) technique if the probabilities of Type 1 and Type 2 classification errors can be estimated. Similarly, domain scores can be adjusted by a Bayesian procedure (e.g., Hambleton & Novick, 1973) to compensate for relative losses associated with the classification errors.

Table 6.9.3

A Comparison of Several Standard Setting Methods

	Judgmental						Combination	
Question	Nedelsky	Nedelsky	Modified Angoff	Angoff	Modified Ebel	Jaeger	Contrasting Groups	Borderline Groups
1. Is a definition of the minimally competent individual necessary?	Yes	Yes	Yes	Yes	Yes	No	No	Yes
2. What is the nature of the rating task—or items, or individuals?	Items	Items	Items	Items	Items	Items	Individuals	Individuals
3. Are examinee data needed?	No	No	No	No	No	No	Yes	Yes
4. Do judges have access to the items?	Yes	Yes	Yes	Yes	Yes	Yes	Usually, but don't need to	Usually
5. Are the judgments made in a group setting or individual setting?	Both	Both	Both	Both	Both	Both	Individual	Individual

When the tests are used for yearly promotions, students' performance in the next grade can be used as a criterion in order to estimate the probabilities of classification errors.

Research is needed on ways of pooling the judgments of several individuals, and of incorporating performance data in primarily content-based judgments.

6.9.4 Professional Licensing/Certification Testing

Tests for licensing and certification differ from the others discussed here in having an external criterion, job performance, which the tests should predict. In addition, these tests are subject to governmental regulations and court rulings on the adequacy with which they reflect requisite job skills (and nothing more). Recent court decisions affirm that content validation of a test against the domain of entry-level job skills is sufficient to demonstrate that the test itself is fair. However, any standard used must also bear a rational relationship to job performance.

One method that will probably be acceptable in the courts is to base the standard on experts' judgments of the importance of each tested item to adequate job performance; that is, to use one of the content-oriented methods to determine a percent correct for passing. The pooled judgments of a large number of expert practitioners would be desirable.

Data on test performance would not be particularly useful in this situation since there is usually not any pre-existing knowledge or belief about the distribution of job-preparedness in the population. Empirical data on criterion (job) performance would be useful were it not for the pervasive selectivity of professions; to use criterion performance properly in establishing optimal passing scores requires an unselected population of job-holders. For these reasons, content-oriented procedures for setting standards are probably the most viable procedures in licensing and certification.

6.10 Summary

In this unit, a number of viable methods for setting standards were introduced. If you wish to view the test by itself and not in relationship to other variables, either Angoff's method or Nedelsky's method appears to be useful. If empirical data is available, Berk's method or the Constrasting Groups method seems especially useful. We have also discussed other methods, of a more complex nature, that are suitable for setting criterion-referenced standards. Our preference for the methods mentioned above stems from the fact that they are simple to implement, and appear to produce defensible results when applied correctly. In the final section of the paper, some proposed sets of procedures for standard setting with respect to three important uses of criterion-referenced tests were outlined. However, considerably more research must be done before these procedures can be recommended for wide-scale use.

We will conclude this unit with a brief discussion of a very important problem. Suppose a set of test items have been selected. If so, it is then possible to set standards via either judgmental or empirical methods (or both). However, if a standard can be set via reference to well-defined domain specifications, and sample test items, tests which will optimally discriminate (i.e., reduce the number of misclassifications) in the region of a standard can be constructed. This is done by selecting test items which "discriminate" in the region of the standard. Test items are piloted on samples of examinees similar to those who will eventually be administered the tests to determine item difficulty levels and discrimination indices. Items with p values near the standard and with the highest discrimination indices are selected for the test. Whether judges can reliably set standards from only domain specifications and some sample test items is unknown. Also, it is not known if standards set by these two different methods will produce different results. This is one of those situations where similar results across two methods would be highly desirable.

6.11 References

Andrew, B. J., & Hecht, J. T. A preliminary investigation of two procedures for setting examination standards. Educational and Psychological Measurement, 1976, 36, 35-50.

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, D.C.: American Council on Education, 1971.

Berk, R. A. Determination of optimal cutting scores in criterion-referenced measurement. Journal of Experimental Education, 1976, 45, 4-9.

Block, J. H. Student learning and the setting of mastery performance standards. Educational Horizons, 1972, 50, 183-190.

Burton, N. Societal standards. Journal of Educational Measurement, 1978, 15, 263-271.

Conaway, L. E. Discussant comments: Setting performance standards based on limited research. Florida Journal of Educational Research, 1976, 18, 35-36.

Conaway, L. E. Setting standards in competency-based education: Some current practices and concerns. Paper presented at the annual meeting of NCME, New York, 1977.

Ebel, R. L. Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall, 1972.

Ebel, R. L. The case for minimum competency testing. Phi Delta Kappan, April, 1978, 546-549.

Educational Testing Service, Report on a study of the use of the National Teachers Examination by the State of South Carolina. Princeton, NJ: Educational Testing Service, 1976.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Glass, G. V. Standards and criteria. Journal of Educational Measurement, 1978, 15, 237-261. (a)

Glass, G. V. Minimum competence and incompetence in Florida. Phi Delta Kappan, 1978, 59, No. 9 (May), 602-605. (b)

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. Journal of Educational Measurement, 1978, 15, 277-290.

Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard-setting. In R. Jaeger & C. Tittle (Eds.), Minimum competency testing.(Approx. Title) Berkeley, CA: McCutchan Publishing Co., 1979.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

Jaeger, R. M. Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 1976, 18, 22-27.

Jaeger, R. M. A proposal for setting a standard on the North Carolina High School Competency Test. Paper presented at the 1978 spring meeting of the North Carolina Association for Research in Education, Chapel Hill, 1978.

Klausmeier, H. J., Rossmiller, R. A., & Saily, M. Individually guided elementary education. New York: Academic Press, 1977.

Kriewall, T. E. Aspects and applications of criterion-referenced tests. Paper presented at the annual meeting of AERA, Chicago, 1972.

Livingston, S. A. A utility-based approach to the evaluation of pass/fail testing decision procedures. Report No. COPA-75-01. Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service, 1975.

Livingston, S. A. Choosing minimum passing scores by stochastic approximation techniques. Report No. COPA-76-02. Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service, 1976.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Nassif, P. M. Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of NCME, Toronto, 1978.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurements. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.

Popham, W. J. Setting performance standards. Los Angeles: Instructional Objectives Exchange, 1978.

Roudabush, G. E. Models for a beginning theory of criterion-referenced tests. Paper presented at the annual meeting of NCME, Chicago, 1974.

Schoon, C. G., Gullion, C. M., & Ferrara, P. Credentialing examinations, Bayesian statistics, and the determination of passing points. Paper presented at the annual meeting of APA, Toronto, 1978.

Shepard, L. A. Setting standards and living with them. Florida Journal of Educational Research, 1976, 18, 23-32.

Torshen, K. P. The mastery approach to competency-based education. New York: Academic Press, 1977.

Van der Linden, W. J., & Mellenbergh, G. J. Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1977, 1, 593-599.

Zieky, M. J., & Livingston, S. A. Manual for setting standards on the Basic Skills Assessment Tests. Princeton, NJ: Educational Testing Service, 1977.

Additional References

Block, J. H. Standards and criteria: A response. Journal of Educational Measurement, 1978, 15, 291-295.

Brennan, R. L., & Lockwood, R. E. A comparison of two cutting score procedures using generalizability theory. ACT Technical Bulletin No. 33. Iowa City, Iowa: American College Testing Program, 1979.

Eignor, D. R. Psychometric and methodological contributions to criterion-referenced testing technology. Unpublished doctoral dissertation, University of Massachusetts, Amherst, 1979.

Emrick, J. A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Levin, H. M. Educational performance standards: Image or substance? Journal of Educational Measurement, 1978, 15, 309-319.

Linn, R. L. Demands, cautions, and suggestions for setting standards. Journal of Educational Measurement, 1978, 15, 301-308.

Popham, W. J. As always, provocative. Journal of Educational Measurement, 1978, 15, 297-300.

Scriven, M. How to anchor standards. Journal of Educational Measurement, 1978, 15, 273-275.

Unit 7

Criterion-Referenced Test and Test Manual Evaluations[1]

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

---

[1] Portions of this unit are from Hambleton, R. K., and Eignor, D. R., Guidelines for evaluating criterion-referenced tests and test manuals. Journal of Educational Measurement, 1978, 15, 321-327.

Table of Contents

Page

7.0 Overview to the Unit. 1

7.1 Introduction. 2

7.2 A Proposed Set of Guidelines. 4

7.3 Evaluation of Eleven Criterion-Referenced Tests 10

7.4 Concluding Remarks. 19

7.5 A State System to Evaluate Criterion-Referenced Tests . . . 20

7.6 References. 37

7.0 Overview to the Unit

The scope and number of criterion-referenced tests available to potential users is impressive. Unfortunately, the quality of these tests varies tremendously and so it is very important for potential users to carefully review available tests before making their selections.

The primary purpose of this unit is to propose a set of guidelines for evaluating criterion-referenced tests and test manuals. The guidelines should be useful to both users and developers of criterion-referenced tests. Secondary purposes are (1) to report on our use of the guidelines with eleven commercially available criterion-referenced test batteries, and (2) to briefly describe a State system to evaluate criterion-referenced tests.

7.1 Introduction

Most of the major test publishers have published in the last few years a wide assortment of criterion-referenced tests. In addition, many school districts, state agencies, small testing firms, and consulting firms have produced their own criterion-referenced tests. Criterion-referenced tests are designed to address many problem areas. For example, criterion-referenced tests are being used to monitor student progress through school programs, to diagnose learning disabilities, to report student progress to parents, to evaluate various types of programs, and to certify or license professionals in many fields. Unfortunately, it appears to us, and to many users of criterion-referenced tests we have spoken with, that many of the available tests fall short of the technical quality necessary for them to accomplish their intended purposes. Perhaps one explanation is that many criterion-referenced tests were developed before an adequate testing technology was fully explicated. Fortunately, there now exists an adequate technology for constructing criterion-referenced tests and using criterion-referenced test scores (Hambleton, Swaminathan, Algina, Coulson, 1978; Popham, 1978). Another possible explanation is that there has been a shortage of guidelines for constructing and using criterion-referenced tests. Certainly the well-known Test Standards for evaluating tests and test manuals prepared by a joint committee of AERA/APA/NCME is helpful, but it is not completely applicable to criterion-referenced tests. Besides the incompleteness of the AERA/APA/NCME Test Standards for evaluating criterion-referenced tests and test manuals, what relevant information there is, is scattered through 75 pages or so of other materials appropriate for norm-referenced test evaluations. Therefore, the Test Standards in its present form, is not very useful for individuals interested in evaluating criterion-referenced tests.

In the next section of this unit, we will propose a set of guidelines for evaluating criterion-referenced tests and test manuals. The guidelines should be useful to both users and developers of criterion-referenced tests. Test standards are not offered (an example of a standard is, "test score reliability must exceed .80"), but we do offer a set of questions for consideration by potential users and developers of criterion-referenced tests. The only other efforts we are aware of to develop guidelines for evaluating criterion-referenced tests and test manuals are Popham (1978, Chapter 8), Swezey and Pearlstein (1975), and Walker (1977). In this unit, we will also report on our use of the guidelines with eleven commercially available criterion-referenced test batteries.

One caution and one comment seem appropriate to introduce at this point. The guidelines represent our own biases about what is important technical information for users to have in making informed decisions about the quality of criterion-referenced tests.

7.2 A Proposed Set of Guidelines

The list of guidelines was generated by placing ourselves in the role of potential purchasers of a criterion-referenced test, and asking "What questions would we want to answer before making a decision to use a criterion-referenced test in a particular situation?" Questions were organized around ten broad categories. They are: Objectives, Test Items, Administration, Test Layout, Reliability, Cut-off Scores, Validity, Norms, Reporting of Test Score Information, and Test Score Interpretations. The questions are as follows:

Objectives

A.1 Is the purpose (or purposes) of the test stated in a clear and concise fashion?

A.2 Is each objective clearly written so that it is possible to identify an "item pool"?

A.3 Is it clear from the list of objectives what the test measures?

A.4 Is an appropriate rationale offered for including each objective in the test?

A.5 Can a potential user "tailor" the test to meet local needs by determining which objectives from a pool of objectives offered by the publisher are to be measured by the test?

A.6 Is there a match between the content measured by the test and the situation where the test is to be used?

A.7 Are individuals identified who were responsible for the preparation of objectives?

A.8 Does the set of objectives measured by the test serve as a representative set from some content domain of interest?

B. Test Items

B.1 Is the item review process described?

B.2 Are the test items valid indicators of the objectives they were developed to measure?

B.3 Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?

B.4 Are the items free of technical flaws?

B.5 Are the test items in an appropriate format to measure the objectives they were developed to measure?

B.6 Are the test items free of bias (for example, sex, ethnic, or racial)?

B.7 Was a heterogeneous sample of examinees employed in piloting the test items?

B.8 Was the item analysis data used only to detect "flawed" items?

C. Administration

C.1 Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?

C.2 Are the test directions clear?

C.3 Is the test easy to score?

C.4 Does the test manual specify an examiner's role and responsibilities?

D. Test Layout

D.1 Is the layout of the test booklets attractive?

D.2 Is the layout of the test booklets convenient for examinees?

E. Reliability

E.1 Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?

E.2 Was the sample (or samples) of examinees used in the reliability study adequate in size, and representative of the population for whom the test is intended?

E.3 Are test lengths suitable to produce tests with desirable levels of test score reliability?

E.4 Is reliability information offered in the test manual for each intended use (or uses) of the test scores?

F. Cut-Off Scores

F.1 Was a rationale offered for the selection of a method for determining cut-off scores?

F.2 Was the procedure for implementing the method explained, and was it appropriate?

F.3 Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?

G. Validity

G.1 Does the validity evidence offered in the test manual address adequately the intended use (or uses) of scores obtained from the test?

G.2 Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?

H. Norms

H.1 Are the norms data reported in an appropriate form?

H.2 Are the samples of examinees utilized in the norming study described?

H.3 Are appropriate cautions introduced for proper test score interpretations?

I. Reporting of Test Score Information

I.1 Are the test scores reported for examinees on an objective by objective basis?

I.2 Are there multiple options available to the user for reporting of test results (for example, by class and grade within a school)?

I.3 Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?

J. Test Score Interpretations

J.1 Are suitable cautions included in the manual for interpreting individual and group objective score information?

J.2 Are appropriate guidelines offered in the manual for utilizing test scores to make descriptive statements, instructional decisions, program evaluation decisions, or other stated uses of the test scores?

A review form, keyed to the 39 guidelines offered above, is presented on the next four pages.

The necessity for many of the guidelines is obvious. For others, brief rationale statements are offered below:

A.4. Rationale statements for the inclusion of particular objectives in a test is especially important in competency-based certification. For example, a manual we saw recently reported that the test, "was designed to measure the skills in reading and mathematics necessary for effective participation in today's complex society." Potential users of the test ought to know the process by which skills or objectives measured by the test were selected or identified.

A.5. Many users desire to have flexibility in the objectives included in their tests.

A.6. Essentially the problem is one of determining content validity. If there is some flexibility in objective selection, it is easier to obtain content valid tests for specific uses.

3/15/79

Criterion-Referenced Test and Test Manual Evaluation Form

Background Information

Test Name: ____________________ Forms and Levels: ____________________

Test Publisher: ____________________ Author(s): ____________________

Year of Publication: ____________________ Cost: ____________________

Reusable Booklets: Yes No Time Limits: ____________________

Special Test Administration Conditions: ____________________

Manual and Other Technical Aids: ____________________

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "√" in the column corresponding to your answer to each question.	Ratings				
Question	Acceptable	Unacceptable	Unsure	Not Applicable	Comments
A.1. Is the purpose (or purposes) of the test stated in a clear and concise fashion?					
A.2. Is each objective clearly written so that it is possible to identify an "item pool"?					
A.3. Is it clear from the list of objectives what the test measures?					
A.4. Is an appropriate rationale offered for including each objective in the test?					
A.5. Can a user "tailor" the test to meet local needs by selecting objectives from a pool of available objectives?					
A.6. Is there a match between the content measured by the test and the situation where the test is to be used?					

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question. Question	Ratings Acceptable	Unacceptable	Unsure	Not Applicable	Comments
A.7. Are individuals identified who were responsible for the preparation of objectives?					
A.8. Does the set of objectives measured by the test serve as a representative set from some content domain of interest?					
B.1. Is the item review process described?					
B.2. Are the test items valid indicators of the objectives they were developed to measure?					
B.3. Is the set of test items measuring an objective representative of the "pool" of items measuring the objective?					
B.4. Are the items free of technical flaws?					
B.5. Are the test items in an appropriate format to measure the objectives they were developed to measure?					
B.6. Are the test items free of bias (for example, sex, ethnic, or racial)?					
B.7. Was a heterogeneous sample of examinees employed in piloting the test items?					
B.8. Was the item analysis data used only to detect "flawed" items?					
C.1. Do the test directions include information relative to test purpose, time limits, practice questions, answer sheets, and scoring?					

BEST COPY AVAILABLE

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question. Question	Ratings				Comments
	Acceptable	Unacceptable	Unsure	Not Applicable	
C.2. Are the test directions clear?					
C.3. Is the test easy to score?					
C.4. Does the test manual specify an examiner's role and responsibilities?					
D.1. Is the layout of the test booklets attractive?					
D.2. Is the layout of the test booklets convenient for examinees?					
E.1. Is the type of reliability information offered in the test manual appropriate for the intended use (or uses) of the scores?					
E.2. Was the sample of examinees adequate in size, and representative of the population for whom the test is intended?					
E.3. Are test lengths suitable to produce tests with desirable levels of test score reliability?					
E.4. Is reliability information offered in the test manual for each intended use (or uses) of the test scores?					
F.1. Was a rationale offered for the selection of a method for determining cut-off scores?					
F.2. Was the procedure for implementing the method explained, and was it appropriate?					

BEST COPY AVAILABLE

For each of the questions below there are four possible answers: "Acceptable", "Unacceptable", "Unsure", and "Not Applicable". Place a "✓" in the column corresponding to your answer to each question.	Ratings				Comments
Question	Acceptable	Unacceptable	Unsure	Not Applicable	
F.3. Was evidence for the validity of the chosen cut-off score (or cut-off scores) offered?					
G.1. Does the validity evidence offered in the test manual address adequately the intended use (or uses of scores) obtained from the test?					
G.2. Is an appropriate discussion of factors affecting the validity of test scores offered in the test manual?					
H.1. Are the norms data reported in an appropriate form?					
H.2. Are the samples of examinees utilized in the norming study described?					
H.3. Are appropriate cautions introduced for proper test score interpretations?					
I.1. Are the test scores reported for examinees on an objective by objective basis?					
I.2. Are there multiple options available to the user for reporting of test results (for example, by class and grade within a school)?					
I.3. Are convenient procedures available for scoring tests by hand, and forms available for reporting test score information?					
J.1. Are suitable cautions included in the manual for interpreting individual and group objective score information?					
J.2. Are appropriate guidelines offered for utilizing test scores to accomplish stated purposes?					

BEST COPY AVAILABLE

A.7. Users ought to know the qualifications and experiences of individuals involved in determining the objectives measured by a test and the process they used in their objectives selection work.

A.8. There appears to be a tendency for some publishers to "slant" their test coverage to objectives easiest to measure. Does the set of objectives measured by the test provide adequate coverage of an area of interest? This is an important question for users to answer.

B.1. Rigorous steps are necessary here. Popham (1978), for example, provides some excellent guidelines that involve many item raters matching items to the objectives the test items were written to measure.

B.2. This can be determined through the use of any one of several rating forms. Face validity evidence is not sufficient.

B.3. The best evidence here is provided by Cronbach's duplication experiment. Alternately, judges can be asked the question directly.

B.4. Standard item writing principles should be used to assess item quality.

E.1. Even when reliability data is reported in a criterion-referenced test manual, it seldom is appropriate for the intended use of the test scores. Standard correlational approaches to reliability provide little relevant information. What is needed, if instructional decisions are to be made, is some indication of the consistency of decision-making over parallel-forms or a retest administration. When the test scores are intended to serve as domain score estimates, some indication of the precision of the estimates should be offered.

E.3. Most users of criterion-referenced tests seem to be unaware of the "large errors" existing in domain score estimates and mastery assignments with short (1 to 5 item) tests.

E.4. Criterion-referenced test scores are used in many ways. Reliability evidence for one use (or in one sample) should not be assumed for other uses (or in other samples).

F.1. There are many methods for setting cut-off scores. A rationale should be offered for any one that is selected. The method should be consistent with the definitions of mastery states offered for sorting examinees.

F.2. Currently there is much debate about setting cut-off scores. At a minimum, to ensure the same value (using the same method) is obtained across different samples of judges, details of the method for determining the cut-off score should be clearly specified.

F.3. The "validity" of a cut-off score can be assessed by relating the classification of examinees based on the particular cut-off score to some independent measure (for example, some outcome measure).

G.1. There are many uses of criterion-referenced test scores. If they are being used for descriptive purposes, evidence of both content and construct validity should be offered. If the test scores are used to sort examinees into mastery states, the relationship between classifications based on the test scores and some appropriately selected independent measure should be reported.

G.2. Again, the problem is no different from that encountered with norm-referenced tests. Since examinees are not being compared with one another, there is a tendency among some publishers to minimize the importance of standardized testing conditions. On the other hand it is becoming more common to prepare norms tables for criterion-referenced tests and therefore standardized test directions in these situations will be important.

H.3. The problem of norms with criterion-referenced tests is about the same as with norm-referenced tests. There is one difference: Criterion-referenced test scores tend to be less reliable because tests are shorter and test scores are often homogeneous. Therefore considerable caution should be used in utilizing normative data. Actually though it is more common to use norms with grouped data, where, fortunately the problem of low individual score reliability is less of a problem.

I.2. Users often desire to have their data summarized in a variety of ways (for example, by class, grade, school, district, sex, race). Are these and other options available?

I.3. When users intend to score their own data it is essential to determine the feasibility of such a strategy. Can the scoring be done conveniently? Are reporting forms available to simplify the process?

J.2. Manuals need to stress the amount of error that exists in criterion-referenced test scores. For example, what is the likelihood of a user making false-positive and false-negative errors? From our experience, we have seldom seen a criterion-referenced test manual that properly cautions test score users about errors in domain score estimation or mastery state determination.

7.3 Evaluation of Eleven Criterion-Referenced Tests

Eleven of the more popular criterion-referenced tests were selected for review. The names of the tests and some descriptive information are presented in Figure 7.3.1.

Our primary purpose was to ascertain the extent to which these tests met our guidelines. We have reported our evaluation of each test relative to each guideline, but the more important information is arrived at by determining how well the tests as a group meet each of our guidelines. The group information is informative because it helps to pin-point areas where commercial materials are in need of revisions and further development.

Figure 7.3.1. Criterion-referenced tests selected for review.

Code	Name of Test	Grades	Levels	Forms	Publication Date	Publisher
1	1976 Stanford Diagnostic Mathematics Test	1-12	4	2	1976	Harcourt Brace Jovanovich
2	1976 Stanford Diagnostic Reading Test	1-12	4	2	1976	Harcourt Brace Jovanovich
3	Skills Monitoring System-Reading	3-5	3	1	1975	Harcourt Brace Jovanovich
4	Individual Pupil Monitoring System-Mathematics	1-6	6	2	1974	Houghton-Mifflin
5	Individual Pupil Monitoring System-Reading	1-8	8	2	1974	Houghton-Mifflin
6	Diagnostic Mathematics Inventory	1.5-7.5	7	1	1977	CTB/McGraw-Hill
7	Prescriptive Reading Inventory	K-6.5	6	1	1977	CTB/McGraw-Hill
8	Diagnosis: An Instructional Aid-Mathematics and Reading	1-6	2	2	1974	Science Research Associates
9	Mastery: An Evaluation Tool-SOBAR Reading	K-9	10	2	1975	Science Research Associates
10	Mastery: An Evaluation Tool-Mathematics	K-8	9	2	1974	Science Research Associates
11	Fountain Valley Support System in Mathematics	K-8	9	1	1974	Richard L. Zweig Associates

In judging the quality of a test and test manual relative to each guideline, the following rating scale was used:

A = Acceptable

A^- = Acceptable, with reservations

X = Unacceptable, data offered was unsuitable or improperly used

Y = Unacceptable, no data was offered

N = Not Applicable

Table 7.3.1 summarizes our ratings of the 11 tests on the 39 guidelines.

Our most significant impressions of the test and test manuals reviewed are as follows:

1. In areas such as Administration, Test Layout, and Norms, there are few problems.

2. Current commercially available "criterion-referenced tests" reviewed in this paper should be called "objectives-referenced tests" since the tests appear to be developed from behavioral objectives (Popham, 1978). Starting to develop a test from a listing of behavioral objectives is less than ideal because behavioral objectives usually do not lead to unambiguous definitions of the "item pools" keyed to the behavioral objectives. The solution is to write "domain specifications" (Popham, 1978).

3. Only about half of the publishers included information about the qualifications of individuals who prepared the objectives measured by their test. The qualifications of participants in this aspect of the test development process is important information for potential users.

Table 7.3.1

Summary of Ratings of the Criterion-Referenced Tests

Question	Test 1	2	3	4	5	6	7	8	9	10	11
A1	A	A	A	A^-	A^-	A	A	A	A	A	X
A2	X	X	X	X	X	X	X	X	X	X	X
A3	A	A	A^-	A	A	A	A	A	A	A	A
A4	A	A	A	A^-	A^-	A	A	A	A	A	X
A5	A^-	A^-	A	A	A	X	X	A	A	A	A
A6	A	A	A	A	A	A	A	A	A	A	A
A7	Y	Y	A^-	Y	Y	Y	A^-	A	A^-	A^-	A^-
A8	A^-	A^-	A^-	A^-	A^-	A^-	A^-	A^-	A^-	A^-	A^-
B1	X	X	A	A^-	A^-	X	A^-	Y	A	A	Y
B2	A^-	A^-	A	A^-	A^-	?[1]	A^-	A^-	A	A	A^-
B3	X	X	X	X	X	X	X	X	X	X	X
B4	A	A	A	A	A	A	A	A	A	A	A
B5	A	A	A	A	A	A	A	A	A	A	A
B6	A	A	A	Y	Y	?	Y	Y	Y	A	Y
B7	A	A	A	A	A	A	A	Y	Y	Y	Y
B8	X	X	A	X	X	X	A^-	Y	X	X	Y
C1	A	A	A	A	A	?	A	A	A	A	?[2]
C2	A	A	A	A	A	?	A	A	A	A	A
C3	A	A	A	A	A	?	A	A	A	A	A
C4	A	A	A	A	A	?	A	A	A	A	A
D1	A	A	A	A	A	?	A	A	A	A	A
D2	A	A	A	A	A	?	A	A	A	A	A
E1	A^-	X	A^-	Y	Y	X	X	Y	X	X	Y
E2	A	A	A	Y	Y	A	A	Y	A	A	Y
E3	A^-	A^-	A^-	A^-	A^-	X	X	X	X	X	A^-
E4	A^-	A^-	A^-	Y	Y	X	X	Y	X	X	Y
F1	A	A	A	Y	A^-	Y	A	X	A	A	Y
F2	A	A	X	Y	Y	X	X	Y	A	A	Y
F3	A	A	A^-	Y	Y	Y	A^-	Y	A^-	A^-	Y
G1	A	A	A	X	X	A	A	X	A^-	A^-	Y
G2	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
H1	A	A	N	N	N	A^-	A	N	N	N	N
H2	A	A	N	N	N	?	Y	N	N	N	N
H3	A	A	N	N	N	Y	Y	N	N	N	N
I1	A	A	A	A	A	?	A	A	A	A	A
I2	A	A	A	A	A	?	A	A	A	A	A
I3	A	A	A	A	A	?	A	A	A	A	A
J1	A^-	A^-	A	Y	Y	?	A^-	Y	A^-	A^-	Y
J2	A	A	A	X	X	?	A	A^-	A^-	A^-	A

---

[1] We did not have the proper materials to assess the quality of the test in the areas marked by a "?".

[2] The information was on a cassette. We did not listen to the tape and so we were not in a position to rate this aspect of the test.

4. Since test developers have not used "domain specifications", it is impossible to assess "item representativeness". Item representativeness is essential if users desire to use objective scores to "generalize to the domains of behaviors defined by the objectives." If item representativeness is not established, scores can only be interpreted in terms of the specific items included in the test.

5. "Item analysis" is an area in which there are two problems: (a) Too little explanation is offered of the choice of particular item statistics and of the specifics of item statistics usage, and (b) item statistics are used in test construction thereby "biasing" the content validity of the test in unknown ways.

6. Test score reliability was not handled very well in most of the manuals. Either (a) inappropriate information relative to the stated uses of the test scores was offered, or (b) no information was offered.

7. Cut-off scores are typically offered, but there is no rationale offered for setting cut-off scores. Procedures used for setting cut-off scores are not explained, nor is any evidence offered for the "validity" of cut-off scores (for example, do those examinees classified as "masters" typically perform better than "non-masters" on some appropriately chosen external criterion measure?).

8. Factors affecting the validity of scores are not offered in any of the manuals.

9. Only a few of the manuals introduced the notion of "error" in test scores. It is extremely important for users to have some indication of the "stability" of their objective scores and/or "consistency of mastery/non-mastery decisions".

7.4 Concluding Remarks

Our proposed guidelines were developed after careful study of the criterion-referenced testing literature and the Test Standards. However, they are offered here only to serve as a "catalyst" for further discussion and debate on a topic of considerable importance to the test and measurement field. Our use of the proposed guidelines to evaluate eleven criterion-referenced tests was intended to (1) demonstrate that the proposed guidelines were workable, and (2) highlight areas where considerably more (or different) work on the part of test developers is needed.

7.5 A State System to Evaluate Criterion-Referenced Tests

George Madaus and Peter Airasian from Boston College and the senior author of the Practitioner's Guidebook prepared a test evaluation system for the Commonwealth of Massachusetts through which content and measurement specialists can determine the appropriateness of commercially available criterion-referenced tests for meeting the Commonwealth's Basic Skills Improvement Policy (BSIP). The BSIP is the Commonwealth's version of a state-wide minimum competency testing program. The program covers the areas of reading and mathematics. School districts must participate in the program in one of three ways: (1) use the Commonwealth's tests, (2) construct and use their own tests, or (3) use one of the commercially available criterion-referenced tests which meet the Commonwealth's content and technical criteria. Our efforts were directed toward the third use. We developed a test evaluation system which includes rating forms, directions for content and technical evaluations, checklists, and summary evaluation sheets. The evaluation system is being used within the Commonwealth to determine which tests meet the Commonwealth's content and technical criteria and therefore can be chosen by school districts for use in complying with the state's minimum competency testing law. On the next few pages are several documents:

1. Standardized achievement test review form
2. Directions for test reviewers — content review
3. Directions for test reviewers — technical review
4. Mathematics skills checklist
5. Standardized achievement test evaluation summary sheet.

The materials are tailored to the BSIP. However, they should help others who have the task of developing test evaluation systems for other states. A report in preparation by Madaus, Airasian and Hambleton will describe the BSIP, steps in the development and validation of the test evaluation system, and several examples of its use.

Standardized Achievement Test

- Review Form[1] -

1. Reviewer ____________________ Date of Review ____________________

3. Test Name ____

4. Test Publisher ____

5. Publication Date ____

6. Levels (Circle Grade Levels Covered by the Test):

K 1 2 3 4 5 6 7 8 9 10 11 12

7. Which form of the test is being reviewed? ____________________

8. Is the test being reviewed for Reading Skills or Math Skills? (Circle one)

Reading Math

If you are doing a content review, begin with Question 9.

If you are doing a technical review, begin with Question 13.

CONTENT CONSIDERATIONS

9. How many of the fourteen reading skills or thirty-eight mathematics skills of the Massachusetts Basic Skills are measured by at least one item on the test? ____ No. of Skills ____ % of Skills

10. Overall, is the reading level of the items reviewed suitable for most of the students in the lowest grade covered by this test? (Cf. Question 6 above). YES NO

---

[1] This review form was prepared by Ron Hambleton, George Madaus and Peter Airasian to meet specifications required by the Commonwealth of Massachusetts for use in conjunction with the Massachusetts Basic Skills Improvement Policy.

11. Overall, are the test items free of offensive sexual, cultural racial, and/or ethnic content and/or stereotyping. YES NO

12. If you answered "NO" to question 11, please explain the reasons for your answer, including the type(s) of bias and the item number of any items of concern.

This is the end of the Content Review

TECHNICAL CONSIDERATIONS

13. How many alternate forms of this test are available? _____ No. of forms

14. Is there a Technical Manual which includes information about the test regarding the following ten topics:

a. Item Review Methods	YES	NO
b. Item Analysis	YES	NO
c. Average Item Difficulty	YES	NO
d. Internal Consistency Reliability	YES	NO
e. Test/Retest Reliability	YES	NO
f. Parallel Form Reliability	YES	NO
g. Standard Error of Measurement	YES	NO
h. Content Validity	YES	NO
i. Norms	YES	NO
j. Procedures for screening items for offensive sexual, cultural, racial, and/or ethnic content, and/or stereotyping	YES	NO

15. How many of the items reviewed meet the standard rules of item writing?

____ No. of items reviewed

____ No. of acceptabl items

____ % of acceptable items

16. Were item analysis results used to identify "defective" test items? YES NO INA*

17. Are data bearing on the consistency of mastery decisions (for one or more performance standards or cut-off scores) reported in the Technical Manual? YES NO

18. Is the consistency of mastery decisions (for one or more cut-off scores) reported in the Technical Manual equal to or above 90%? YES NO INA

19. Do standard indices of internal consistency reliability reported on the total reading score or total mathematics score reach or exceed .90? YES NO INA

20. Do standard indices of test-retest or parallel form reliability as reported on the total reading score or total mathematics score reach or exceed .90? YES NO INA

21. If parallel-forms of the Test are available, do both forms (or multiple-forms, if available) measure equally well the content spanned by the skills included in the Test? (In other words, do the multiple-forms of the Test have equivalent content validity?) YES NO INA

22. Are the test score norms based on data that is no more than five years old? YES NO INA

23. Were the norm groups of sufficient size (i.e., at least 300 students)? YES NO INA

24. Were the samples of students used in the norming study representative of students in the grades for which this test is intended? (Cf. Question 6) YES NO INA

25. Were the samples of students used in the norming study representative of important strata within the society (i.e., rural pupils, minority group pupils, pupils in large city schools, etc.) YES NO INA

---

*INA - Information not available

26. Are the test administration directions suitable for students in the lowest grade covered by the test? (Cf. Question 6) YES NO

If "NO", please explain ______________________________

____

____

____

27. Do the test administration directions address the matter of time limits? YES NO

If "NO", please explain ______________________________

____

____

____

28. Do the test administration directions indicate to the student how to handle the problem of guessing? YES NO

If "NO", please explain ______________________________

____

____

____

29. Is the layout or format of the test booklet convenient for students in the lowest grade covered by the test? (cf Question 6) YES NO

If "NO", please explain ______________________________

____

____

____

30. Is the layout or format of the answer sheet convenient for students in the lowest grade covered by the test? (Cf. Question 6) YES NO

If "NO", please explain ______________________________

____

____

____

31. Does the test include practice questions? YES NO

This is the end of the Technical review

Directions for Test Reviewers
- Content Review -

The content review you are about to undertake involves three principal tasks:

a. Deciding whether each of the test items the publisher has nominated as measuring each of the fourteen reading skills or thirty-eight mathematics skills of the Massachusetts Basic Skills Policy in fact is appropriate indicators of the skill in question.

b. Deciding whether overall the reading level of the items on the test is suitable for the majority of students in the lowest grade covered by the test.

c. Deciding whether overall the test is free of offensive sexual, cultural, racial or ethnic content and/or stereotyping.

You are asked to make a determination on each of these points by completing the enclosed Review form. Three people will review each test and will meet to arrive at a composite rating for each test. A separate technical review of each test is also being carried out.

To begin the review you should have the following materials in front of you:

a. A copy of the reading or math tests to be reviewed.

b. A list of the test items which the test publisher feels correspond to each of the fourteen reading skills or thirty-eight mathematics skills of the Massachusetts Basic Skills Policy.

c. A skills checklist which lists the fourteen reading skills (blue color) or thirty-eight math skills (yellow color).

d. A Standardized Achievement Test Review Form.

e. A Standardized Achievement Test Evaluation Summary Sheet (pink color).

Step A. - Complete the "Basic Information" section of the Standardized Achievement Test Review Form (Questions 1 - 8).

Fill out the background information section on the Skills Checklist and on the Test Evaluation Summary Sheet.

Step B. - Read carefully through the list of skills included in the Skills Checklist.

Read carefully through all the test items on the reading or mathematics test under review.

Step C. - Question 9 on the Review Form

For each skill listed on the Skills Checklist read each item which the publisher has nominated as a measure of that skill. If you agree that the item is a valid indicator of the skill in question, list the item number in the space provided. Once you have finished with a skill, count up the number of items nominated by the publisher which you feel are valid indicators of the skills and place the total number in the blank space provided on the Skills Checklist.

If at least one item nominated by the publisher is a valid indicator of the skill in question you should place a "✓" beside the Commonwealth's skill listed on the Skills Checklist in the box provided.

After you have completed your review of each of the nominated questions for each of the fourteen reading skills or thirty-eight mathematics skills, add up the total number of acceptable items across all the skills and place your total in the space provided at the end of the check list. Next in the space provided write the total number of items on the reading or math test reviewed.

Finally count up the number of "✓" marks (i.e., each skill that has at least one item you feel is a valid indicator of that skill). Place the total number of "✓" in the space provided in Question 9 on the Review Form. Calculate the percent of skills measured by at least one test item. For example, suppose 8 of the Commonwealth's 14 reading skills are measured by at least one item on a Test. You would write "57" in the space provided beside Question 9 for percent of skills included in the test.

Step D. - Question 10 on the Review Form

This item is self-explanatory. Make your decision on the basis of your reading of all the items on the test. For example if the test is designed for 7th, 8th, and 9th graders (indicated in Question 6) the reading level should be appropriate for 7th graders.

Step E. - Questions 11 and 12

Question 11 - After reading through all the items on the test, decide whether overall the test is free of offensive sexual, cultural, racial, and/or ethnic content and/or stereotyping. You should examine all test items to determine whether there is a consistent or overriding pattern of racial, ethnic, cultural, or sexual stereotyping and/or offensive content. Your judgment should be made within the context of the total test. The fact that one or two items portray a woman in the kitchen or a minority group member in an unskilled occupation does not necessarily imply stereotyping. Some women do spend time in the kitchen and some minority group members do hold unskilled jobs. At issue is whether mem-of such groups are consistently or predominantly portrayed in such circumstances relative to the way in which other groups are portrayed.

Question 12 - Self-explanatory.

Step F. - Transfer the information from the Review Form to the Test Evaluation Summary Sheet.

Thank you your time and effort.

Directions for Test Reviewers

- Technical Review -

The technical review you are about to undertake involves making judgments about certain technical characteristics of tests which are being considered for possible inclusion on a State-approved list of standardized commercial tests. Local school districts may use a test on the list to assess basic skills in reading and mathematics at the secondary level (grades 7-12).

Three people will review each test and will meet to arrive at a composite rating for each test. A separate content review of each test is also being carried out to assess the test's content validity relative to the Massachusetts Basic Skills Policy.

To begin the review you should have the following materials in front of you:

a. Copies of the test to be reviewed.

b. Copies of the Technical Manual for each test.

c. A Standardized Achievement Test Review Form.

d. A Standardized Achievement Test Evaluation Summary Sheet (pink color).

Step A - Complete the "Basic Information" section of the Standardized Achievement Test Review Form, Questions 1 - 8.

Fill out the background information section on the Test Evaluation Summary Sheet.

Step B - Read carefully through the test booklets and the Test's Technical Manual.

Step C - THE TECHNICAL REVIEW BEGINS AT QUESTION 13. Complete each of the following questions on the Review Form:

Questions 13 and 14 - Self-explanatory

Question 15 - Read the technical aid, "Multiple-Choice Item Writing Principles" on page 32, and then randomly select and review 25% of the test items to determine the percent of these test items which do not violate any of the standard rules of multiple-choice item writing. Write the number of items reviewed, the number of acceptable items and the percent of item reviewed which are acceptable in the spaces provided beside Question 15 on the review form.

Question 16 - Check to be sure that item difficulties and item discrimination indices were used in any item analyses. (In constructing criterion-referenced tests, however, the latter is a more important and useful statistic.

INA means Information Not Available.

Questions 17 and 18 - Check for the proportion of agreement in decision-making across parallel-form or retest administrations. Alternately, check to see if the statistic, k, is reported. It reflects the proportion of agreement over and above agreement which is due to chance alone.

Questions 19 and 20 - The test manual will most likely report numerous reliability indices. In general, do these indices reach or exceed .90?

Question 21 - Check to see if the content validity of two (or more) forms is the same. Often the Technical Manual will discuss content emphases and summarize the relevant information in charts or tables. If this information is not satisfactory the parallel forms will be reviewed separately another time by another review committee.

Questions 22 and 23 - Self-explanatory.

Questions 24 and 25 - Check to see if charts are produced to show the representation of any norms groups. Do they look reasonable?

Questions 26 to 31 - These five questions are self-explanatory.

Step D - Transfer the information from the Review Form to the Test Evaluation Summary Sheet.

THANK YOU FOR YOUR TIME AND EFFORT

BEST COPY AVAILABLE

Multiple-Choice Item Writing Principles

1. Is the item stem clearly written for the intended group of students?
2. Is the item stem free of irrelevant material?
3. Is a single problem clearly defined in the item stem?
4. Are the answer choices clearly written for the intended group of students?
5. Are the answer choices free of irrelevant material?
6. Is there a correct answer or a clearly best answer?
7. Have words like "always," "none," or "all" been removed?
8. Are likely student mistakes used to prepare incorrect answers?
9. Is "all of the above" avoided as an answer choice?
10. Are the answer choices arranged in a logical sequence (if one exists)?
11. Was the correct answer randomly positioned among the available answer choices?
12. Are all repetitious words or expressions removed from the answer choices and included in the item stem?
13. Are all of the answer choices of approximately the same length?
14. Do the item stem and answer choices follow standard rules of punctuation and grammar?
15. Are all negatives underlined?
16. Are grammatical cues between the item stem and the answer choices, which might give the correct answer away, removed?
17. Are letters used in front of the possible answer choices to identify them?
18. Have expressions like "which of the following is not" been avoided?

- Mathematics Skills Checklist[1] -

Reviewer ________________________________ Date of Review __________

Test Name ____

Place a "✓" beside those skills which are measured by the test.

Mathematics Skills

a. Number and Numeration Concepts

1. Recognize number symbols (17, eighteen), whole numbers (34), fractions (1/2), decimals (3.75), and powers of 10 (10^2). ☐

   List the number of each item which you feel is a measure of this skill.

   ____

   ____ ☐

   Total number of items for this skill ____________________

2. Identify odd and even numbers. ☐

   List the number of each item which you feel is a measure of this skill.

   ____

   ____

   Total number of items for this skill ____________________

3. Put numbers in numerical order. ☐

   List the number of each item which you feel is a measure of this skill.

   ____

   ____

   Total number of items for this skill ____________________

---

[1] Only the first page of the mathematics skills checklist is presented here.

Standardized Achievement Test
Evaluation Summary Sheet

Reviewer ______________________________ Date of Review ______________

Test Name ____

Check one – Reading _________ Math _________

Fill in your ratings, determine the points, and write in the score for each question in the space provided.

CONTENT CONSIDERATIONS

Question	Rating	Point System	Score
9	___%	90-100%-5 points 80- 89%-4 points 70- 79%-3 points 60- 69%-1 point < 60%-0 points	___
10	___	Yes - 2 points No - 0 points	___
11	___	Yes - 3 points No - 0 points	___
12		No points	

TOTAL CONTENT POINTS . . . . . ☐

TECHNICAL CONSIDERATIONS

Question	Rating	Point System	Score
13		No points	
14	a___	Yes - 1 point	a___
	b___	No - 0 points	b___
	c___	for each item	c___
	d___	"a" through "j"	d___
	e___		e___
	f___		f___
	g___		g___
	h___		h___
	i___		i___
	j___		j___

Question	Rating	Point System	Score
15	___%	90-100%-5 points 80- 89%-4 points 70- 79%-3 points < 70%-0 points	___
16	___	Yes - 3 points No - 0 points INA*- 0 points	___
17	___	Yes - 1 point No - 0 points	___
18	___	Yes - 5 points No - 0 points INA - 0 points	___
19	___	Yes - 5 points .80-.89-3 points .70-.79-1 point less than .70-0 points INA - 0 points	___
20	___	Yes - 5 points .80-.89-3 points .70-.79-1 point less than .70-0 points INA - 0 points	
21	___	No points However if No or INA then alternative forms of the test are subject to a separate review at another time	
22	___	Yes - 2 points No - 0 points INA - 0 points	___
23	___	Yes - 2 points No - 0 points INA - 0 points	___

---

*INA - "Information not available"

Question	Rating	Point System	Score
24	___	Yes - 3 points No - 0 points INA - 0 points	___
25	___	Yes - 3 points No - 0 points INA - 0 points	___
26	___	Yes - 2 points No - 0 points INA - 0 points	___
27	___	Yes - 2 points No - 0 points	___
28	___	Yes - 2 points No - 0 points	___
29	___	Yes - 2 points No - 0 points	___
30	___	Yes - 2 points No - 0 points	___
31	___	Yes - 2 points No - 0 points	___

TOTAL TECHNICAL POINTS . . . . . ☐

7.6 References

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, NJ: Prentice-Hall, 1978.

Swezey, R. W., & Pearlstein, R. B. Guidebook for developing criterion-referenced tests. A report prepared for the U.S. Army Research Institute for the Behavioral and Social Sciences. Reston, VA: Applied Science Associates, August, 1975.

Walker, C. B. Standards for evaluating criterion-referenced tests. Los Angeles: Center for the Study of Evaluation, UCLA, 1977. (Unpublished manuscript.)

Unit 8

Using and Reporting Test Score Information

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

Table of Contents

Page

8.0 Overview of the Unit. 1

8.1 Introduction. 2

8.2 Uses of Criterion-Referenced Test Scores. 3

8.3 Domain Score Estimation 5

8.3.1 Introduction 5
*8.3.2 Specialized and Bayesian Estimates 6

8.4 Mastery State Determination 15

8.4.1 Introduction 17
*8.4.2 Advanced Decision Models 25

*8.5 Simulation Study Involving Criterion-Referenced Test Scores 38

8.6 Reporting of Information. 47

8.6.1 Individual Student 47
8.6.2 Group. 50

8.7 Grading 63

8.8 References. 68

Note: Starred ("*") sections may be omitted without loss of continuity. These sections involve the use of Bayesian statistical methods with criterion-referenced test scores.

8.0 Overview of the Unit

Procedures for using and reporting criterion-referenced test scores are discussed in this unit.

8.1 Introduction

The first five units of the Practitioner's Guidebook covered methods for developing and validating criterion-referenced tests. In Unit 6, issues and methods for setting standards were introduced. A list of guidelines for evaluating criterion-referenced tests and test manuals was presented in Unit 7. Ways in which test scores obtained through applications of criterion-referenced tests can be used and reported will be considered in this unit. First, we discuss several uses of criterion-referenced test scores. Second, we will discuss and provide examples of the ways in which criterion-referenced test score information can be reported.

In sum, in this unit, we hope to give the reader some practical information on ways to use and/or report criterion-referenced test score information. In the next unit, we will extend our discussion presented here to the application of criterion-referenced tests in two popular instructional models.

8.2 Uses of Criterion-Referenced Test Scores

Millman (1974) delineates four uses of criterion-referenced test scores:

1. estimation of domain scores and allocation to mastery states in instructional settings,

2. evaluation of programs,
3. needs assessment purposes,
4. teaching improvement and personnel evaluation.

The focus of the material to be discussed in the remainder of Unit 8 will be on the first use listed. Hambleton and Gifford have prepared a paper that discusses the use of criterion-referenced tests in program evaluation. (After final editing, this paper will be included in these instructional materials.) Besides Millman (1974), Popham (1975) also has a discussion of the uses of criterion-referenced tests for program evaluation.

In reference to program evaluation, Millman (1974) states:

> One consideration in the evaluation of instructional programs is the degree to which the objectives of the program have been met. DRT's [domain referenced tests] are designed to present such information. Further, in contrast to national, norm-referenced tests, tests referencing the specific domains of learner behaviors to which the instructional effort is directed have a better chance of detecting areas in which the program has been successful or is in need of modification.

In reference to the use of criterion-referenced tests in assessing needs, it is helpful to first discuss the meaning of need. According to Millman (1974), "A need can be defined as the difference between expected and actual status." In other words, a discrepancy exists between present status and what is expected. However, before movement can be initiated for change, the present status of the area to be changed (for instance, an instructional program) must be determined. A criterion-referenced test is most useful in establishing present status.

In reference to teaching improvement, Millman (1974) states:

> When student performance is measured by DRT's [domain referenced tests], the desired student behavior becomes explicit. The precise boundaries of the behavior to be assessed are defined, and criteria for judging the adequacy of learner responses are identified. Such information makes it possible for the teacher to devise more relevant instructional materials and provides for a fairer evaluation of the teacher's performance.

In the sections to follow, we will focus attention on two uses of criterion-referenced test scores in instructional settings. These uses are (1) estimation of examinee domain scores, and (2) allocation of examinees to mastery states.

8.3 Domain Score Estimation

8.3.1 Introduction

In this section, the basic problems involved with domain score estimation are introduced. Then we will discuss specialized procedures for estimating domain scores (section 8.3.2). The discussion in 8.3.2 has been taken from a paper by Hambleton, Swaminathan, Algina and Coulson (1978).

We assume that a test is constructed by randomly sampling items from a well-defined, or clearly specified, domain of items measuring an instructional objective. If the test measures more than a single objective, the items must be randomly sampled from the domain of items measuring each objective. (An examinee has a domain score defined for each objective measured by the test.)

In problems of domain score estimation, it is common to use an examinee's test score (or proportion-correct score) as an estimate of the domain score of that examinee. An examinee's domain score is his/her true level of performance in the domain of items measuring the objective. Of course, there will be error involved in using the observed test score as an estimate of the domain score, and that is why specialized methods involving Bayesian procedures have been developed. The estimates derived from these procedures are more precise (i.e., contain less error) estimates of the domain score.

When using the test score, or some other derived (i.e., Bayesian) estimate, to estimate an examinee's domain score, error can be defined as the difference between the estimated and true value (i.e., domain score). The test developer thus wants to choose as his/her estimate the one that minimizes these differences over the group of individuals tested.

As mentioned above, the simplest and most obvious estimate of an examinee's domain score, which is denoted π_i for the ith examinee, is his/her observed proportion-correct score, denoted $\hat{\pi}_i$. This estimate is obtained by dividing the examinee's test score, x_i (the number of items answered correctly), by the total number, n, of items measuring the objective included in the criterion-referenced test. Although the proportion-correct score is an unbiased estimate of domain score, this estimate is highly unreliable when the number of items on which the estimate is based is small. For this reason, specialized procedures that take into account other available information in order to produce more precise estimates, especially when there are only a few items on the test measuring an objective, are used. In section 8.3.2, we discuss a number of such estimates.

*8.3.2 Specialized or Bayesian Estimates of Domain Score

The estimates discussed in this section utilize additional information besides an individual's proportion-correct score to arrive at examinee domain score estimates. However, to obtain these estimates requires the use of a small computer to carry out the somewhat complicated calculations.

Classical Model II Estimate

One of the first attempts to produce an estimate of an examinee's true score using the information obtained from the group to which an examinee belongs was made by Kelley in 1927. This is the well-known regression estimate of true score (Lord and Novick, 1968, p. 65), which is the weighted sum of two components — one based on the examaminee's observed score and the other based on the mean of the group to which an examinee belongs. Jackson (1972) modified this procedure for use with binary data, by employing the Freeman-Tukey transformation, given by

$$g_i = \frac{1}{2} \left(\sin^{-1} \sqrt{\frac{x_i}{n+1}} + \sin^{-1} \sqrt{\frac{x_i+1}{n+1}} \right) . \qquad (1)$$

As a result of this transformation, a domain score is transformed onto γ_i, where,

$$\gamma_i = \sin^{-1} \sqrt{\pi_i} . \qquad (2)$$

When, $.15 \leq \pi \leq .85$, and the number of test items (n) is at least eight, the distribution of g_i is approximately normal with a mean approximately equal to the transformed domain score, γ_i, and known variance

$$v = (4n + 2)^{-1} .$$

The classical model II estimate becomes, in terms of γ,

$$\hat{\gamma}_i = [g_i \hat{\phi} + (4n + 2)^{-1} g.] \; / \; [\hat{\phi} + (4n + 2)^{-1}] , \qquad (3)$$

where g., the sample mean based on a sample of N examinees, is given by

$$g. = N^{-1} \sum_{i=1}^{N} g_i , \qquad (4)$$

and $\hat{\phi}$, the sample variance of the γ's, is given by

$$\hat{\phi} = (N - 1)^{-1} \sum_{i=1}^{N} (g_i - g.)^2 - (4n + 2)^{-1} . \qquad (5)$$

Once $\hat{\gamma}_i$ is obtained, $\hat{\pi}_i$ is determined from the expression

$$\hat{\pi}_i = (1 + .5/n) \sin^2 \hat{\gamma}_i - .25/n. \qquad (6)$$

For a detailed discussion of this estimate, the reader is referred to Novick and Jackson (1974, p. 352) and Novick, Lewis, & Jackson (1973).

Bayesian Model II Estimate

The classical model II estimate given above may not be ideal since it does not take into account any prior information that may be available. In addition, it may happen that $\hat{\phi}$ estimated using Equation (5) is negative, in which case the solution will not be meaningful. Novick et al. (1973) utilizing the transformations given by Equations (1) and (2), obtained a Bayesian solution for the estimation of domain score that not only takes into account the direct and collateral information, but also any prior information that may be available. Direct information is provided by an examinee's test score; collateral information is contained in the test performance of other examinees; prior information on an examinee may come from past test performance or the examinee's performance on other objectives measured by the test. In addition, the Bayesian model II estimation procedure avoids the problem of negative estimates for ϕ.

Since the distribution of g_i has known variance but unknown mean γ_i, the distribution of g_i is customarily expressed as a conditional distribution, i.e.,

$$(g_i \mid \gamma_i) \sim N(\gamma_i, v) \tag{7}$$

where $N(\gamma_i, v)$ represents the normal distribution with mean γ_i and variance v. The Bayesian estimates are based on the revised belief about the parameters after the data are obtained. The revised belief about the parameters after the data are obtained is summarized in the form of the posterior distribution of the parameters.

In order to obtain the posterior distribution of γ_i, it is necessary to specify the prior knowledge about the distribution of γ_i, or $f(\gamma_1, \gamma_2, \ldots, \gamma_N)$. In order to do this, it is assumed that the transformed domain scores $\gamma_1, \gamma_2, \ldots, \gamma_N$ are exchangeable. This amounts to saying that the prior belief about one γ_i is no different from the belief about any other γ_j and implies the assumption that γ_i is a random sample from some distribution. In particular, it is assumed that the prior distribution of γ_i is normal with unknown mean α and unknown variance ϕ. Thus, the specification of the prior distribution of γ_i is dependent upon the knowledge of the mean α and the variance ϕ. However, Novick, et al. (1973) have suggested that the prior belief about α may not be as important as the specification of the prior belief about ϕ and may be represented by a uniform distribution. The above authors have further assumed that it is reasonable to represent the belief about ϕ by an inverse chi-square distribution with ν degrees of freedom and scale parameter λ (see Novick and Jackson, 1974, for an extensive discussion of this distribution). Specification of the prior belief about ϕ thus requires the specification of only the two parameters, ν and λ.

Novick, et al. (1973) have considered in detail the problem of setting values of the parameters, ν and λ. Based on various considerations, these authors recommend setting $\nu = 8$. The mean $\bar{\phi}$, of the inverse chi-square distribution is given by $\lambda / (\nu-2)$, and once ν is known, λ can be set equal to $(\nu-2)\,\bar{\phi}$. To estimate $\bar{\phi}$ it is necessary to indicate the amount of information that is available about π. This is accomplished by specifying a value M, where M is considered to be the π value of the typical examinee in the sample. The next step is to specify the number of test items, n, that would have to be administered to the examinee in order to obtain as much information about π as is deemed to be available. Now, transformed estimates of π, from a n-item test are distributed normally on the γ-metric with variance $(4n + 2)^{-1}$. Hence, $(4n + 2)^{-1}$ can be taken as an estimate of $\bar{\phi}$ and subsequently λ can be specified.

Specification of ν and λ in essence determines the prior distribution $f(\gamma)$ of $\gamma_1, \gamma_2, \ldots, \gamma_N$.
Novick, et al. (1973) obtained the joint posterior distribution of the parameters, and hence the joint modal estimate of γ_i.

The joint modal estimate γ_i is obtained by solving the equation

$$\gamma_i = \frac{g_i\left[\frac{\lambda + \Sigma(\gamma_i - \gamma.)^2}{N + \nu - 1}\right] + \gamma.\left[\frac{1}{(4n + 2)}\right]}{\left[\frac{\lambda + \Sigma(\gamma_i - \gamma.)^2}{N + \nu - 1}\right] + \left[\frac{1}{(4n + 2)}\right]} \quad . \tag{8}$$

where

$$\gamma. = N^{-1} \sum_{i=1}^{N} \gamma_i \quad . \tag{9}$$

This equation for γ_i has to be solved iteratively, and has been found (Novick, et al. 1973) to yield a satisfactory solution after only a few iterations.

Marginal Mean Estimate

The Bayesian model II estimate discussed above is useful for making joint decisions about a set of N examinees. However, in criterion-referenced testing situations, a separate decision for each examinee has to be made and hence separate or marginal estimates of domain scores are required.

Lewis, Wang, and Novick (1973) obtained marginal mean estimates of domain scores. They are given by the expression

$$\hat{\gamma}_i = g. + \rho^*(g_i - g.) \quad . \tag{10}$$

The quantity ρ^* is dependent on the parameters ν and λ and on the data; once the parameters are set, ρ^* can be read directly from tables prepared by Wang (1973). Again, once $\hat{\gamma}_i$ is obtained $\hat{\pi}_i$ is determined using Equation (6).

The marginal mean estimate given above is based on the assumption that no prior information is available on α, i.e., the prior distribution of α is uniform. More recently, Lewis, Wang and Novick (1975) relaxed this assumption by assuming that α is normally distributed with mean θ and variance ϕ/η. In this case, they showed that the estimate $\hat{\gamma}_i$ is given by

$$\hat{\gamma}_i = \rho g_i + (\tau-\rho) g_. + (1-\tau)\,\theta \ . \qquad (11)$$

Since the definitions of ρ and τ are rather involved, we refer the interested reader to Lewis, et al. (1975) for a discussion of these quantities and for the procedure required to specify the additional parameters, θ and η .

"Quasi" Bayesian Estimates

In obtaining the joint modal estimates and the marginal mean estimates, Novick, et al. (1973) and Lewis, et al. (1973) assumed that the prior beliefs about α and ϕ could be expressed in the form of distributions. There are several variations to this theme. If instead of specifying the prior beliefs in the form of distributions, values for α and ϕ can be specified on the basis of previous experience, then the expressions corresponding to the Bayesian marginal mean estimates are readily obtained. These estimates are relatively easy to compute.

These estimates are based on the prior specification of α and ϕ. Specification of α introduces relatively few complications, but the exact specification of ϕ poses a problem. This is not a quantity most practitioners are familiar with. However, the interrogation procedure described by Novick and Jackson (1974) can be effectively used to yield this information. Two assumptions are made in deriving these quasi-Bayesian estimates: (1) The prior belief about α can be expressed as a uniform distribution, and ϕ can be specified exactly, and, (2) both α and ϕ can be specified exactly. In the first case, it can be shown that the marginal mean estimate $\hat{\gamma}_i$ is given by

$$\hat{\gamma}_i = \frac{g_i \phi + (4n+2)^{-1} g.}{\phi + (4n+2)^{-1}} . \qquad (12a)$$

In the second case, the marginal mean estimate, $\hat{\gamma}_i$, becomes

$$\hat{\gamma}_i = \frac{g_i \phi + (4n+2)^{-1} \alpha}{\phi + (4n+2)^{-1}} . \qquad (12b)$$

The similarity between the marginal mean estimates (12a) and (12b) and the classical model II estimate given by Equation (3) is obvious. In fact, it is interesting to note that the classical model II estimate is in reality an empirical Bayes estimate obtained by using sample estimates for α and ϕ.

Hambleton, Hutten, and Swaminathan (1976) investigated the comparative efficiencies of the various estimates given by Equations 3, 8, 10, 12a, 12b, and the proportion-correct score estimate via a simulation study (see section 8.5). Factors under consideration in their study were sample size, test length, homogeneity of the domain score distribution, specification of prior information and cut-off score (or performance standard). Their conclusions indicated that when precise information is available, i.e., when ϕ and α can be specified, the marginal mean estimate, $\hat{\gamma}_i$, given by Equation (12b), had the smallest absolute error as defined by the expression

$$e = \sum_{i=1}^{N} \left| \gamma_i - \hat{\gamma}_i \right| .$$

When α cannot be specified exactly, the estimate given by Equation (12a) produced the next best result in terms of minimizing e. The other estimates, ranging from third best to poorest were: Marginal mean estimate given by Equation (10), classical model II estimate given by Equation (3), the joint modal estimate given by Equation (8), and the proportion-correct score estimate. However, in most cases α and ϕ cannot be specified exactly, and hence, the results of this study bear out the expectation that Bayesian estimation procedures are the most efficient in the estimation of domain scores. Also, it should be pointed out that these authors did not study the estimate given by Equation (11). We can, nevertheless, conclude that the estimate given by Equation (11) would be at least as accurate as that given by Equation (10) if the assumption of a normal prior on α is valid.

8.4 Mastery State Determination

In this section, the basic situation involved when a criterion-referenced test score is being used to allocate an individual to a mastery state is introduced. Then, as in section 8.3, some advanced decision models and a Bayesian procedure for making examinee assignments to mastery states are discussed. The material is presented in section 8.4.2.

Before discussing advanced procedures for allocating examinees to mastery states, it will be useful to review a section first encountered in Unit 4 : Types of errors (false positive and false negative) made when classifying individuals into mastery states. The following two-fold table of losses associated with decisions can be constructed:

		Domain Scores	
		$\pi \geq \pi_o$	$\pi < \pi_o$
Decision	Advance	o	a
	Retain	b	o

Where π = the examinee's domain score

a = loss associated with advancing a student whose domain score π is $<\pi_o$ (false positive error)

b = loss associated with retaining a student whose domain score π is $\geq\pi_o$ (false negative error).

The values of a and b are specified by the test constructor. One possible decision is to let $a = b$, and this might be done, according to Novick and Lewis (1974), when

> . . .it were no more serious to advance a student whose level was below the criterion than to retain a student who was above, ...

From the specification above, a general decision rule can be generated. The rule is to advance (assign to a "mastery state") an examinee if

$$b[\text{Prob}(\pi\geq\pi_o|\text{data})] > a[\text{Prob}\ (\pi<\pi_o|\text{data})]$$

and retain (i.e., assign to a "non-mastery state"), otherwise. An equivalent comparison is to compare the loss ratio $\frac{a}{b}$ to the ratio

$$\frac{\text{Prob}\ (\pi\geq\pi_o|\text{data})}{\text{Prob}\ (\pi<\pi_o|\text{data})}$$

If $\frac{a}{b}$ is less than the above ratio, the examinee should be advanced. If $\frac{a}{b}$ is greater than the above ratio, the examinee should be retained. Prob $(\pi\geq\pi_o)$ is the probability of an examinee having a domain score equal to or above the cut-off score. The probability is obtained as a part of a Bayesian analysis of the examinee's test performance.

8.4.1 Introduction

A second major use of scores obtained from criterion-referenced tests is to assign examinees to mastery states. In view of the discussion in section 8.3, it may appear tempting to first estimate an examinee's domain score, compare it to one or more cut-off scores defined on the domain score scale, and then, for example, in the case of two mastery states, classify the examinee as either a master or a non-master. Typically, this is the strategy adopted by individuals implementing objectives-based instructional programs. Unfortunately, this approach is not usually very satisfactory. One reason is that users must assume all classification errors (whether they be of the "false-positive" or "false-negative" type) to be equally serious (i.e., a = b). This is an unreasonable assumption to make in many instructional settings. For example, with instructional objectives that are prerequisites to more advanced ones in a curriculum, false-positive errors (moving examinees ahead before they are ready) may be far more serious than false-negative errors (holding examinees back, even though they may have "mastered" the objectives in question). (One possible solution is to raise a cut-off score when false-positive errors are more serious than false-negative errors. When the importance of the errors is reversed, a cut-off score can be lowered.) Also, domain score estimates may be obtained using a loss function completely inappropriate for that associated with making decisions. In assigning an examinee to a mastery state, an error can occur when an examinee is assigned, based upon his/her test score (or a variation of it), to a mastery state other than his/her true mastery state. For example, the individual may truly have mastered the material, but based upon his/her score on the test, be assigned non-mastery status.

Here,the notion of error used in domain score estimation (squared error loss) makes no sense. Distance or difference, in this case from the relevant cut-off score, is not a concern; rather the concern is simply whether the examinee located either above or below the cut-off score is correctly assigned to the proper mastery state. Thus, the appropriate loss function in this decision-theoretic process would be a threshold loss function. On the other hand, Livingston (1972, 1975), and Linden and Mellenbergh (1977) have investigated both linear and non-linear loss functions. Here, one assumes the misclassification of an examinee with a domain score far from the cut-off score is far more serious than the loss incurred when an examinee with a domain score close to a cut-off score is misclassified.

For the test practitioner who lacks the facilities to enable him/her to use the somewhat complex methods that follow, then he/she should determine mastery status by comparing an individual's proportion correct score to the cut-off score. However, such a procedure suffers the same problems discussed in section 8.3.1. If the number of items measuring an objective is small, then the proportion correct score will often give an unreliable estimate for determining mastery status. Also, and perhaps more of a problem, when using this procedure, all classification errors must be assumed to be equal. That is why the procedures to be discussed is section 8.4.2 are so valuable; they incorporate additional data into the estimates, thereby decreasing the error, and they allow for the consideration of different classification errors.

Throughout this introduction, we have been implicitly assuming the existence of only two mastery states, master and non-master. However, in some decision contexts, there may be more than two states. For instance, there may be two cut-offs, π_r and π_a, such that if the student's score is below π_r, he/she is retained. If his/her score is above π_a, he/she is advanced, and finally, if the score is between the two cut-off scores, the individual may be "held" for a short review. In the development that follows, the procedures are first formulated for one cut-off point (two mastery states), and then extended to k mastery states.

The problem of classifying an examinee into one of two categories using a threshold loss function has been studied extensively by Hambleton and Novick (1973), and Swaminathan, Hambleton, and Algina (1975). As in section 8.3.2, the observed scores x_i are transformed into g_i by the arc sine transformation. Let π and π_o denote the domain score and cut-off score respectively, and $\gamma(=\sin^{-1}\sqrt{\pi})$ denote the transformed domain score π, and γ_o $(=\sin^{-1}\sqrt{\pi_o})$ the transformed cut-off score. Then, examinees with transformed domain scores γ less than γ_o are classified as non-masters and masters otherwise. Conforming with the notation employed by Hambleton and Novick (1973), the two-valued parameter ω is used to denote the mastery state of an examinee. The parameter ω assumes one of two values, ω_1 or ω_2. If the examinee is a non-master, i.e., if $\gamma < \gamma_o$, then

$$\omega = \omega_1,$$

while if the examinee is a master, i.e., $\gamma \geq \gamma_o$,

$$\omega = \omega_2 .$$

In classifying an examinee the decision-maker may take one of two actions - for example, retain the examinee for instruction or advance the examinee to the next segment of instruction. The action "retain" will be denoted by a_1 and the action "advance" by a_2. The decision-maker can commit one of two kinds of errors. If the examinee is in reality a non-master (in state ω_1), the decision-maker can classify the examinee as a master (in state ω_2) or if in reality the examinee is a master (in state ω_2), the decision-maker can classify the examinee as a non-master (in state ω_1). In order to arrive at a rule for selecting actions a_1 or a_2, it is necessary to specify the losses associated with these two kinds of classification errors.

Swaminathan, et al. (1975), introduced the quantity, $L(\omega_i, a_j)$, to denote the non-negative loss function describing the loss incurred when action a_j is taken for an examinee who is in state ω_i. Thus, in the two category decision problem,

$$L(\omega_1, a_2) = \ell_{12},$$

and

$$L(\omega_2, a_1) = \ell_{21} .$$

with

$$L(\omega_1, a_1) = L(\omega_2, a_2) = 0 .$$

These authors have suggested that the action for which the expected loss

$$E_{\omega}L(\omega, a)$$

is a minimum should be chosen as the appropriate action.

BEST COPY AVAILABLE

Swaminathan, et al. (1975) extended the two category problem to one where examinees are classified into one of several categories. Suppose there are k categories into which the examinees are to be classified and consequently k actions to be taken. For example, when k=3, the decision-maker may be interested in classifying examinees as masters, partial masters, and non-masters. The appropriate actions may be to advance the masters, retain the partial masters for a brief review, and retain the non-masters for remedial work.

We need k-1 cut-off scores to separate examinees into k categories or k states, $\omega_1, \omega_2, \ldots, \omega_k$. Denote the cut-off scores by $\pi_{o1}, \pi_{o2}, \ldots, \pi_{ok-1}$. An examinee is in state ω_1, when her domain score π is less than π_{o1}, in state ω_2 if π is between π_{o1} and π_{o2}, and so on. In general an examinee is in state ω_i if $\pi_{oi-1} \leq \pi < \pi_{oi}$.

Associated with misclassifications is the loss function $L(\omega_i, a_j)$. If an action a_j is taken for an examinee who in reality is in state ω_i, the loss is ℓ_{ij} so that

$$L(\omega_i, a_j) = \ell_{ij} .$$

As before, the action which has the smallest expected loss is chosen.

For action a_j, the expected loss is given by

$$E_\omega L(\omega, a_j) = \sum_{p=1}^{k} \ell_{pj} \text{ Prob } [\gamma_{op-1} \leq \gamma < \gamma_{op} | \text{Data}] \tag{13}$$

where $\gamma_{oo} = -\infty$, and $\gamma_{ok} = +\infty$. Action a_j is chosen if

$$\sum_{p=1}^{k} \ell_{pj} \text{ Prob } [\gamma_{op-1} \leq \gamma < \gamma_{op} | \text{Data}]$$

$$< \sum_{p=1}^{k} \ell_{pm} \text{ Prob } [\gamma_{op-1} \leq \gamma < \gamma_{op} | \text{Data}], \quad (m=1, 2, \ldots, k, m \neq j). \tag{14}$$

BEST COPY AVAILABLE

Once the posterior distribution of γ is determined, the above probabilities are determined as the area under the probability density curve between γ_{op-1} and γ_{op}, $p = 1, 2, \ldots, k$.

The next stage in the decision-theoretic process is to obtain the posterior distribution of parameter, γ, for each examinee. Several procedures are available for the determination of posterior distributions and, hence, posterior probabilities. The first method is that given by Lewis, et al. (1973). Utilizing the distributions and assumptions given in connection with the Bayesian Model II estimates in a previous section, Lewis, et al. (1973) derived an approximation to the posterior distribution. They showed that the posterior distribution of γ_i, is approximately normal, i.e.,

$$(\gamma_i \mid \text{Data}) \sim N(\mu_i, \sigma_i^2) \tag{15}$$

where

$$\mu_i = g. + \rho^*(g_i - g.), \tag{16}$$

and

$$\sigma_i^2 = \frac{1 + (N - 1)\,\rho^*}{(4n + 2)\,N} + (g_i - g.)^2\,\sigma^{*2}. \tag{17}$$

(This approximation is reasonably good when the number of test items

BEST COPY AVAILABLE

exceeds seven.) The quantity $g_.$ is defined by Equation (4). The quantities ρ^* and σ^{*2} in expressions (16) and (17) are dependent on the parameters ν and λ of the inverse chi-square distribution of ϕ, and have to be computed by numerical integration.
The tables prepared by Wang (1973) can be used by specifying ν and λ, to obtain ρ^* and σ^{*2}.

Returning to the problem of classification of examinees into k categories, Swaminathan et al. (1975) first transform the (k-1) specified cut-off scores π_{op} into γ_{op}, given by

$$\gamma_{op} = \sin^{-1}\sqrt{\pi_{op}}, \qquad p = 1,\ldots,k-1. \qquad (18)$$

Next the probabilities of the type given by Equations (13) and (14) are calculated. For any examinee,

$$\text{Prob}[\pi_{op-1} \leq \pi < \pi_{op} \mid \text{Data}] = \text{Prob}[\gamma_{op-1} \leq \gamma < \gamma_{op} \mid \text{Data}]. \qquad (19)$$

For the ith examinee, the quantity z_{oji} is defined as

$$z_{oji} = \frac{\gamma_{oj} - \mu_i}{\sigma_i}, \qquad (20)$$

with μ_i and σ_i^2 defined by Equations (16) and (17). The quantity z_{oji} is the normal deviate corresponding to the cut-off score γ_{oj} for examinee i. Since the posterior distribution is approximately normal with mean μ_i and variance σ_i^2,

$$\text{Prob}[\gamma_{op-1} \leq \gamma_i < \gamma_{op} \mid \text{Data}] = \text{Prob}[z_{op-1i} \leq z_i < z_{opi} \mid \text{Data}]. \qquad (21)$$

That is, the probability that γ_i is between γ_{op-1} and γ_{op} is approximately equal to the probability that a standardized normal variate is between the z scores z_{op-1} and z_{op}. Hence, for each examinee i, the quantity

$$E_\omega L(\omega,a_j) = \sum_{p=1}^{k} \ell_{pj} \, \text{Prob}[z_{op-1i} \leq z_i < z_{opi} \mid \text{Data}] \qquad (22)$$

is calculated for each action j (j=1, 2,...,k). These k expected losses are than compared with one another, and the action for which the expected loss is the least is chosen as the appropriate action. An illustration of the procedure is offered in the next section.

*8.4.2 A Bayesian Decision Theoretic Procedure

The paper by Swaminathan, Hambleton, and Algina (1975) describes one method for using Bayesian decision-theoretic procedures to allocate examinees to mastery states.

JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 12, NO. 2 SUMMER 1975

A BAYESIAN DECISION-THEORETIC PROCEDURE FOR USE WITH CRITERION-REFERENCED TESTS[1]

H. SWAMINATHAN, RONALD K. HAMBLETON, and JAMES ALGINA
University of Massachusetts

In a previous paper, Hambleton and Novick (1973) conceptualized a decision-theoretic formulation for several issues in criterion-referenced measurement. Among the issues discussed was the important problem of allocating individuals to mastery states. These authors proposed a solution to the problem based on a Bayesian procedure given by Novick, Lewis, and Jackson (1973). More recently, Lewis, Wang, and Novick (1973) have developed a Bayesian procedure that is more appropriate in the context of criterion-referenced measurement. Based on this most recent method, we present in this paper an exposition of a decision-theoretic solution to the problem of allocating individuals to mastery states on the objectives included in a criterion-referenced test.

Allocation of Individuals to Mastery States

The primary problem in criterion-referenced measurement is that of classifying an examinee into one of several mutually exclusive mastery states or categories. One might think of mastery states, defined for an objective, as representing different levels of functioning on the domain of items measuring that objective. It makes sense to assume that each examinee has a true mastery state on each objective covered in a criterion-referenced test. Typically, a cut-off score or threshold score is set to permit the decision-maker to assign examinees, on the basis of their performance on each subset of items measuring an objective covered in a criterion-referenced test, into one of two mutually exclusive categories—masters and non-masters. (See, Millman, 1973, for a good discussion of guidelines for setting cutting scores.) Since all the items in the domain of items measuring an objective cannot usually be administered to the examinee, a small number of items is sampled. Thus the problem of classifying the examinees into categories has to be considered within a statistical framework.

An obvious approach to the allocation problem is to compare an examinee's observed score to the threshold score and make the appropriate mastery decision. However, as criterion-referenced tests are typically short, we would be making decisions on the basis of very limited amounts of information. Our decision-theoretic approach to the allocation problem allows the decision-maker to build into the decision process prior and collateral information about the examinee's *true* mastery state. This approach is not unlike that of using a regression line to estimate true scores, and provides a way of obtaining more information on each examinee without requiring the administration of additional test items—a great advantage indeed when one considers the amount of time typically allotted for criterion-referenced testing in objectives-based programs (see, for example, Glaser & Nitko, 1971 pp. 625-670; Hambleton, 1974; Hambleton & Novick, 1973). However, even after incorporating this additional information, our knowledge of an examinee's true mastery state will be probabilistic and misclassifications will be likely to occur. The decision-theoretic approach also allows the decision-maker to incorporate into the decision process the costs of misclassifications.

---

[1] The authors are grateful to Ming-Mei Wang, Douglas Coulson, and Jason Millman for helpful comments on earlier drafts of the paper.

Classification of Examinees Into One of Two Categories

We shall first consider the problem of classifying an examinee into one of two categories and later generalize the procedure to include several categories.

Let γ denote the "true" score of an examinee. We will see later that γ is related in a very simple way to π, the true proportion-correct score. The quantity π is defined as the proportion of items, in the domain of items measuring the objective, that an examinee can correctly answer. If γ_o is the predetermined threshold or cut-off score, examinees with true scores γ less than γ_o are classified as true non-masters and true masters otherwise. In keeping with the notation employed by Hambleton and Novick (1973) let the two-valued parameter ω denote the mastery state of the examinee. The parameter ω assumes one of two values, ω_1 or ω_2. If the examinee is a non-master, i.e., if $\gamma < \gamma_o$, we set

$$\omega = \omega_1,$$

and if he is a master, i.e., $\gamma \geq \gamma_o$, we set

$$\omega = \omega_2.$$

Both γ and ω are, of course, unobservable quantities. Our approach is to produce, using Bayesian statistical methods, a distribution representing our belief about the location of the parameter γ. Using this distribution, known as the posterior distribution on the true score parameter, γ, and with a cutting score defined, we can produce probabilities representing the chances of an examinee being located in each mastery state.

In classifying the examinees, the decision-maker may take one of two actions—retain the examinee for instruction or advance the examinee to the next segment of instruction. The action "retain" will be denoted by a_1 and the action "advance" by a_2. The decision-maker can commit one of two kinds of errors. If the individual is in reality a non-master (in state ω_1), the decision-maker can classify the individual as a master (in state ω_2) or if in reality the individual is a master (in state ω_2), the decision-maker can classify the individual as a non-master (in state ω_1). In order to arrive at a rule for selecting actions a_1 or a_2, it is necessary to specify the losses associated with these two kinds of misclassifications.

Consistent with the usage and notation of decision theory, we shall employ the notation $L(\omega_i, a_j)$ to denote the non-negative loss function which describes the loss incurred when action a_j is taken for the individual who is in state ω_i. Thus,

$$L(\omega_1, a_2) = \ell_{12},$$

and

$$L(\omega_2, a_1) = \ell_{21}.$$

Of course,

$$L(\omega_1, a_1) = L(\omega_2, a_2) = 0.$$

A good classification procedure is obviously one which minimizes in some sense or other the total loss incurred. That is, we shall choose that action for which the expected loss,

$$E_\omega L(\omega, a),$$

is a minimum.

We see that if action a_1 is taken, then the expected loss, $E_\omega L(\omega, a_1)$, is given by

$$\begin{aligned} E_\omega L(\omega, a_1) &= 0 \cdot \mathrm{Prob}[\omega = \omega_1] + \ell_{21}\,\mathrm{Prob}[\omega = \omega_2] \\ &= \ell_{21}\,\mathrm{Prob}[\gamma \geq \gamma_o]. \end{aligned} \tag{1a}$$

Similarly, if action a_2 is taken, then the expected loss, $E_\omega L(\omega, a_2)$, is given by

$$\begin{aligned} E_\omega L(\omega, a_2) &= \ell_{12}\,\mathrm{Prob}[\omega = \omega_1] + 0 \cdot \mathrm{Prob}[\omega = \omega_2] \\ &= \ell_{12}\,\mathrm{Prob}[\gamma < \gamma_o]. \end{aligned} \tag{1b}$$

We take action a_1 if

$$E_\omega L(\omega, a_1) < E_\omega L(\omega, a_2),$$

or equivalently, if

$$\ell_{21}\,\mathrm{Prob}[\gamma \geq \gamma_o] < \ell_{12}\,\mathrm{Prob}[\gamma < \gamma_o]. \tag{2a}$$

Similarly, we take action a_2 if

$$\ell_{12}\,\mathrm{Prob}[\gamma < \gamma_o] < \ell_{21}\,\mathrm{Prob}[\gamma \geq \gamma_o]. \tag{2b}$$

If it so happened that

$$\ell_{12}\,\mathrm{Prob}[\gamma < \gamma_o] = \ell_{21}\,\mathrm{Prob}[\gamma \geq \gamma_o],$$

we would be indifferent as to which action to take.

In order to clarify the meaning of $\mathrm{Prob}[\gamma < \gamma_o]$ and $\mathrm{Prob}[\gamma \geq \gamma_o]$, and hence the expected losses given by (2a) and (2b), we have to distinguish between *prior* probabilities and *posterior* probabilities. In simplistic terms, prior probabilities are based on our beliefs about the parameter γ before any test data are obtained. For example, we often have information about the ability levels of the students in a program in the form of school records, their past performance, etc. This information, conveniently summarized in the form of the prior distribution $f(\gamma)$ of the parameter γ, reflects our prior belief before new test data are obtained. The posterior probabilities on the other hand, are based on our revised belief about the parameter γ after the test data are obtained. And this belief is summarized by the posterior distribution of the parameter γ, denoted by say, $h(\gamma \mid \mathrm{Data})$. In the language of statistics, $h(\gamma \mid \mathrm{Data})$, the posterior distribution of γ, is the conditional distribution of γ given the data. The area under the curve $h(\gamma \mid \mathrm{Data})$ below γ_o gives the probability, $\mathrm{Prob}[\gamma < \gamma_o \mid \mathrm{Data}]$, and the area above γ_o gives the probability, $\mathrm{Prob}[\gamma \geq \gamma_o \mid \mathrm{Data}]$.

Unfortunately, the posterior distribution for each examinee is not obtainable directly. The first stage in obtaining this posterior marginal distribution is to obtain the joint posterior distribution of all the m examinees, $h(\gamma_1, \gamma_2, \ldots, \gamma_m \mid \mathrm{Data})$.

As a consequence of Bayes Theorem, the posterior joint distribution is readily expressed in terms of the joint prior distribution $f(\gamma_1, \gamma_2, \ldots, \gamma_m)$ as

$$h(\gamma_1, \gamma_2, \ldots, \gamma_m \mid \mathrm{Data}) \propto g(\mathrm{Data} \mid \gamma_1, \gamma_2, \ldots, \gamma_m) f(\gamma_1, \gamma_2, \ldots, \gamma_m). \tag{3}$$

The expression $g(\text{Data} \mid \gamma_1, \gamma_2, \ldots, \gamma_m)$ is known as the likelihood function and is a statement of the joint probability of observing the data conditioned upon the unknown parameters $\gamma_1, \gamma_2, \ldots, \gamma_m$. We shall return to the discussion of obtaining the posterior marginal distribution from the joint posterior distribution in the next section.

The probabilities in expressions (2a) and (2b) are, in actuality, posterior probabilities and hence should be so denoted. Thus, we take action a_1 if

$$\ell_{21} \text{Prob}[\gamma \geq \gamma_o \mid \text{Data}] < \ell_{12} \text{Prob}[\gamma < \gamma_o \mid \text{Data}], \tag{4a}$$

and take action a_2 if

$$\ell_{12} \text{Prob}[\gamma < \gamma_o \mid \text{Data}] < \ell_{21} \text{Prob}[\gamma \geq \gamma_o \mid \text{Data}]. \tag{4b}$$

Description of the Bayesian Decision-Theoretic Procedure

We begin by assuming that the ith examinee is administered a random sample of n items measuring a particular objective. An examinee has a true proportion-correct score, π, defined over the domain of items measuring the objective. Although it would be possible to obtain an estimate of π on the basis of the examinee's performance on the sample of test items, this is not our primary aim. If we consider testing within a decision-making framework, then to make decisions concerning an examinee's mastery state, we need the posterior probabilities of the kind described by Equations (4a) and (4b).

Since it is mathematically inconvenient to work with π, we shall, following Novick et al. (1973), utilize the transformation

$$\gamma = \sin^{-1}\sqrt{\pi} \tag{5}$$

and obtain the posterior distribution of γ instead of π. To be compatible with this transformation all of our observed test scores will need to be transformed to the γ-metric. This is easily accomplished by transforming the test score x_i of the ith examinee into

$$g_i = \sin^{-1}\{(x_i + 3/8)/(n + 3/4)\}^{1/2}. \tag{6}$$

This particular transformation, which has been discussed in some detail by Novick et al. (1973), is attractive because, for examinee i, the distribution of g_i is approximately normal with mean $\gamma_i = \sin^{-1}\sqrt{\pi_i}$ and constant variance $V = (4n + 2)^{-1}$. The approximation is reasonably good when π lies between .15 and .85 and n is at least 8. Since the distribution of g_i has known variance but unknown mean γ_i, the distribution of g_i is customarily expressed as a conditional distribution, i.e.,

$$g_i \mid \gamma_i \sim N(\gamma_i, V) \tag{7}$$

where $N(\gamma_i, V)$ represents the normal distribution with mean γ_i and variance V. Referring to Equation (3) we can see that in order to obtain the posterior distribution for each γ_i, we need the likelihood function $g(\text{Data} \mid \gamma_1, \gamma_2, \ldots, \gamma_m)$. The product of the m distributions given by Equation (7), where m is the number of examinees in the sample, yields the likelihood function.

In order to obtain the posterior distribution of γ_i, we have to specify our prior knowledge about the distribution of γ_i. We assume that the transformed "true" scores, $\gamma_1, \gamma_2, \ldots, \gamma_m$, of the m individuals are exchangeable. This amounts to saying that our prior belief about one γ_i is no different than our belief about any other γ_j and implies the assumption that each γ_i is randomly sampled from some distribution. In particular, we assume that the prior distribution of γ_i is normal with unknown mean α and unknown variance ϕ. Thus, the specification of the prior distribution of γ_i is dependent upon our knowledge of the mean α and the variance ϕ. However, Novick et al. (1973) suggest that our prior belief about α may not be as important as our prior belief about ϕ. The above authors have assumed that it is reasonable to represent our belief about ϕ by an inverse chi-square distribution with ν degrees of freedom and scale parameter λ (see Novick & Jackson, 1974, for an extensive discussion of this distribution). Specification of the prior belief about ϕ thus requires the specification of only the two parameters, ν and λ.

Novick et al. (1973) have considered in detail the problems of setting values of the parameters, ν and λ. Based on theoretical considerations, these authors recommend setting $\nu = 8$. The mean $\bar{\phi}$, of the inverse chi-square distribution is given by $\lambda / (\nu - 2)$, and once ν is known, λ can be set equal to $(\nu - 2)\,\bar{\phi}$. To estimate $\bar{\phi}$ we are required to indicate the amount of information we have about π. This is accomplished by specifying a value M, where M is considered to be the π value of the typical examinee in the sample. We then specify the number of test items, t, we would need to administer to the examinee in order to obtain as much information about π as we feel we now have. Transformed estimates of π, from a t-item test are distributed normally on the γ-metric with variance $(4t + 2)^{-1}$. Hence, we could take $(4t + 2)^{-1}$ as our estimate of $\bar{\phi}$ and subsequently specify λ.

Specification of ν and λ in essence determines the prior distribution $f(\gamma)$ of $\gamma_1, \gamma_2, \ldots, \gamma_m$. Substituting this in Equation (3), Novick et al. (1973) obtained the joint posterior distribution of the parameters. This joint posterior distribution of $\gamma_1, \gamma_2, \ldots, \gamma_m$ is useful for making joint decisions about the m individuals. However, in criterion-referenced testing situations we are interested in making separate decisions about each individual and hence we require the distribution of each γ_i; i.e., the marginal distribution of γ_i.

It has been shown by Lewis et al. (1973) that the posterior marginal distribution of γ_i, our belief about the location of the ith examinee's score on the γ-metric, is approximately normal, i.e.,

$$\gamma_i \mid \text{Data} \sim N(\mu_i, \sigma_i^2) \tag{8}$$

where

$$\mu_i = \bar{g} + \rho^*(g_i - \bar{g}). \tag{9}$$

and

$$\sigma_i^2 = \frac{1 + (m - 1)\rho^*}{(4n + 2)m} + (g_i - \bar{g})^2 \sigma^{*2}. \tag{10}$$

(This approximation is reasonably good when the number of test items exceeds seven.) The quantity $\bar{g}$, is defined as

$$\bar{g} = m^{-1} \sum_{i=1}^{m} g_i.$$

The quantities ρ^* and σ^{*2} in Equations (9) and (10) are dependent on the parameters

ν and λ of the inverse chi-square distribution of ϕ, and have to be computed by numerical integration. Wang (1973) has prepared a set of tables so that on specifying ν and λ, ρ^* and σ^{*2} may be obtained.

Returning to the problem of classification of students into masters and nonmasters, we first transform the specified cut-off score π_o into γ_o, given by

$$\gamma_o = \sin^{-1} \sqrt{\pi_o}. \tag{11}$$

Now we have to calculate the probabilities necessary for comparisons of the type given by Equations (4a) and (4b). For any individual it should be clear that

$$\text{Prob}[\pi_i \geq \pi_o \mid \text{Data}] = \text{Prob}[\gamma_i \geq \gamma_o \mid \text{Data}].$$

For each individual we define the quantity z_{oi} as

$$z_{oi} = \frac{\gamma_o - \mu_i}{\sigma_i}, \tag{12}$$

with μ_i and σ_i^2 defined by Equations (9) and (10). Since the posterior distribution is approximately normal with mean μ_i and variance σ_i^2,

$$\text{Prob}[\gamma_i \geq \gamma_o \mid \text{Data}] \simeq \text{Prob}[z \geq z_{oi} \mid \text{Data}].$$

That is, the probability that γ_i is greater than γ_o is approximately equal to the probability that a standardized normal variate is greater than the z score, z_{oi}. Hence

$$\ell_{12}\, \text{Prob}[z < z_{oi} \mid \text{Data}]$$

can be compared with

$$\ell_{21}\, \text{Prob}[z \geq z_{oi} \mid \text{Data}]$$

and the appropriate decision made.

For convenience we shall summarize the procedure by outlining the steps taken to arrive at the appropriate action and illustrate the procedure with a hypothetical example.[2] The steps are:

1. Transform the number correct x_i for the ith examinee into g_i, given by

$$g_i = \sin^{-1} [(x_i + 3/8)/(n + 3/4)]^{1/2}.$$

2. Specify the cut-off score π_o and obtain the corresponding γ_o, given by

$$\gamma_o = \sin^{-1} \sqrt{\pi_o}.$$

3. Specify the prior distribution of ϕ by specifying the parameters ν and λ.
4. Obtain ρ^* and σ^{*2} as tabulated by Wang (1973) and hence determine the mean μ_i and variance σ_i^2 of the posterior distribution of γ_i given by equations (9) and 10.
5. Obtain the standardized normal deviate

$$z_{oi} = (\gamma_o - \mu_i)/\sigma_i$$

---

[2] For another description of the steps we refer the reader to the excellent document on criterion-referenced measurement prepared by Millman (1974, pp. 311-397).

and hence determine the probability, Prob[$z \geq z_{oi}$ | Data], which is approximately equal to Prob[$\pi_i \geq \pi_o$ | Data].

6. Make the decision according to Equations (4a) or (4b).

We will illustrate the above procedure by the following hypothetical example. Our data and results are summarized in Tables 1 and 2. Suppose that a set of 10 items representative of the domain of items measuring an objective is administered to a group of 25 examinees, and that the cut-off score π_o is set at .80. First, we transform the observed scores, x_i into g_i, and the cut-off score π_o into γ_o. Next, we must specify our prior beliefs about ϕ. As indicated earlier, we do this by choosing ν and λ, the parameters of the distribution that we use to represent our belief about ϕ. In order to determine ν and λ, we must decide the length of the test that would be required to give us as much information as we feel we now have about any examinee's true mastery score π_i. Suppose that, in our example, we decided that a five-item test would be required. We therefore take $t = 5$ and, hence, $(4t + 2)^{-1} = .0455$, as our value for $\bar{\phi}$. Since, in general, a good value for ν is eight, the value for λ is .2730, because $\lambda = (\nu - 2)\,\bar{\phi}$. Using the tables prepared by Wang (1973), we find $\rho^* = .5335$ and $\sigma^{*2} = .0159$. We now have enough information to compute μ_i and σ_i using Equations (9) and (10). Finally, we obtain the standardized normal deviate given by Equation (12) and using the tables of the standardized normal distribution find the approximate probability, Prob[$\pi_i \geq .8$ | Data] and its complement Prob[$\pi_i < .8$ | Data]. Suppose also that the loss associated with a false-positive error, ℓ_{12}, is taken to be one "unit" and the loss associated with a false-negative error, ℓ_{21}, to be two "units." In order to make a decision about each examinee we weight the appropriate probability by the associated loss and obtain the expected loss for each action. Thus,

Table 1

Bayesian Analysis of a Hypothetical Set of Data: n=10, m=25

Number of Items Correct x_i	Frequency	Transformed Observed Score g_i	Marginal Mean μ_i	Marginal Standard Deviation σ_i
4	2	.695	.836	.121
5	4	.785	.881	.118
6	5	.875	.933	.118
7	4	.980	.989	.115
8	4	1.083	1.043	.115
9	3	1.202	1.107	.118
10	3	1.392	1.211	.125

$\bar{g} = m^{-1}\Sigma g_i = .998$

BEST COPY AVAILABLE

Table 2

Decision-Making in the Two-Category Classification Problem: n=10, m=25, ℓ_{12}=1, ℓ_{21}=2

Number of Items Correct x_i	Prob[π_i<.8|Data]	Prob[$\pi_i \geq$.8|Data]	Expected losses		Action
			Action a_1 (Retain) =ℓ_{21} Prob[$\pi_i \geq$.8|Data]	Action a_2 (Advance) =ℓ_{12} Prob[π_i<.8|Data]	
4	.988	.012	.025	.988	retain
5	.972	.028	.056	.972	retain
6	.931	.069	.139	.931	retain
7	.849	.151	.302	.849	retain
8	.710	.290	.579	.710	retain
9	.502	.498	.994	.502	advance
10	.231	.770	1.539	.231	advance

$\gamma_o = \sin^{-1} \sqrt{\pi_o} = 1.107$

taking the losses into account, examinees with nine or ten correct items are advanced, while examinees with less than nine correct items are retained for instruction. By manipulating the various losses in the example it is easy to see how other decisions may be made.

Classification of Examinees Into One of k Categories

Suppose that there are k categories into which the examinees are to be classified, and consequently, k actions to be taken. For example, when $k = 3$, the decision-maker may be interested in classifying examinees as masters, partial masters, or non-masters. The appropriate actions may be to advance the masters, retain the partial masters for a brief review and send the non-masters for remedial work.

In order to separate examinees into k categories or k states, $\omega_1, \omega_2, \ldots, \omega_k$, we need k-1 cut-off scores. Denote these by $\pi_{o1}, \pi_{o2}, \ldots, \pi_{ok-1}$. Hence, an examinee is in state ω_1, if his true proportion-correct score π is less than π_{o1}, in state ω_2 if his score π is between π_{o1} and π_{o2}, and so on. In general an examinee is in state ω_i if $\pi_{oi-1} \leq \pi < \pi_{oi}$. In addition, we denote the set of k actions by $a_1, a_2, \ldots, a_j, \ldots, a_k$. Action a_j is to be taken if the examinee is classified into state ω_j.

Associated with misclassifications is the loss function $L(\omega_i, a_j)$. If an action a_j is taken for an individual who in reality is in state ω_i, the loss is ℓ_{ij} so that

$$L(\omega_i, a_j) = \ell_{ij}.$$

These losses are conveniently displayed in Table 3. As before, we choose the action which has the smallest expected loss. Here again we utilize the transformation presented in Equation (5).

For action a_j, the expected loss is given by

Table 3
Loss Table for a
Multi-Action Problem

State	Action					
	a_1	a_2	...	a_j	...	a_k
ω_1 $(\gamma < \gamma_{o1})$	0	ℓ_{12}	...	ℓ_{1j}	...	ℓ_{1k}
ω_2 $(\gamma_{o1} \leq \gamma < \gamma_{o2})$	ℓ_{21}	0	...	ℓ_{2j}	...	ℓ_{2k}
...	...	...	...	...	...	...
ω_i $(\gamma_{oi-1} \leq \gamma < \gamma_{oi})$	ℓ_{i1}	ℓ_{i2}	...	ℓ_{ij}	...	ℓ_{ik}
...	...	...	...	...	...	...
ω_k $(\gamma_{ok-1} \leq \gamma)$	ℓ_{k1}	ℓ_{k2}	...	ℓ_{kj}	...	0

$$E_\omega L(\omega, a_j) = \sum_{i=1}^{k} \ell_{ij} \operatorname{Prob}[\gamma_{oi-1} \leq \gamma < \gamma_{oi} \mid \text{Data}],$$

where $\gamma_{oo} = -\infty$, and $\gamma_{ok} = +\infty$. Thus action a_j is chosen if

$$\sum_{i=1}^{k} \ell_{ij} \operatorname{Prob}[\gamma_{oi-1} \leq \gamma < \gamma_{oi} \mid \text{Data}] <$$
$$\sum_{i=1}^{k} \ell_{ip} \operatorname{Prob}[\gamma_{oi-1} \leq \gamma < \gamma_{oi} \mid \text{Data}] \quad (p = 1, 2, \ldots, k, p \neq j).$$

The probability $\operatorname{Prob}[\gamma_{oi-1} \leq \gamma < \gamma_{oi} \mid \text{Data}]$ is calculated in the manner described in the last section.

In order to illustrate the procedure in the multiple action problem, we utilize the hypothetical data given in Table 1. Suppose that the losses associated with wrongly classifying an examinee into one of three categories, masters, partial masters, and non-masters, are as reported in Table 4.

Assuming that the cutting scores, π_{o1} and π_{o2}, are .65 and .80 respectively, and working with the posterior distribution of γ for each examinee in exactly the same manner as in the previous example, it is possible to calculate the probability of each examinee being in any of the three mastery states. The hypothetical probabilities reported in Table 5 are the probabilities associated with an examinee being in any of

Table 4
Hypothetical Losses for the Three-Action Problem

State	Action a_1 (Remedial Work)	 a_2 (Brief Review)	 a_3 (Advance)
Non-Master	0	2	3
Partial Master	1	0	2
Master	2	1	0

these three categories. These probabilities, when combined with the loss structure presented in Table 4, would result in examinees with six or fewer correct items being retained for remedial work, examinees with seven, eight, or nine correct items being retained for a brief review, and examinees with ten items correct being moved ahead.

Table 5

Decision-Making in the Three-Category Classification Problem: n=10, m=25

Number of Items Correct	Prob[$\pi_i<.6$|Data]	Prob[$.6\le\pi_i<.8$|Data]	Prob[$\pi_i\ge.8$|Data]	Expected Losses Action a_1	 Action a_2	 Action a_3	Action
4	.659	.329	.012	.353	1.330	2.635	retain
5	.516	.456	.028	.512	1.060	2.460	retain
6	.345	.586	.069	.724	.759	2.207	retain
7	.184	.665	.151	.967	.519	1.882	retain briefly
8	.087	.623	.290	1.203	.464	1.507	retain briefly
9	.031	.471	.498	1.467	.560	1.035	retain briefly
10	.005	.225	.770	1.765	.780	.665	advance

$\gamma_{o1} = \sin^{-1}\sqrt{.6} = .886$

$\gamma_{o2} = \sin^{-1}\sqrt{.8} = 1.107$

Conclusion

The procedure described in this paper should be feasible with objectives-based programs that have a small computer of the type typically used to manage instruction (see, for example, Baker, 1971). We shall attempt to demonstrate the feasibility of the procedure by briefly outlining the steps a hypothetical instructional designer would

BEST COPY AVAILABLE

take. Let us suppose that an instructional designer is interested in making decisions on students' status with respect to a particular set of program objectives. Test items measuring each objective are organized into a criterion-referenced test and administered to the students. We assume that the test items are binary scored and represent a random sample of items from the domain of items that measure each objective. For each objective, the designer must specify the number and the location of the mastery states on the mastery score interval [0, 1]. This is done by defining the cutting scores. In addition, the instructional designer specifies the losses attached to classifying an individual incorrectly. A loss matrix of the kind shown in Table 3 is developed and provided to the computer. Some rough guidelines for developing the loss matrix have been described by Hambleton and Novick (1973). Finally, it is necessary for the designer to specify his prior beliefs about the distribution of ability on each objective covered in the test. This is one step where the instructional designer needs to be extremely careful. The effects of poor choice of priors on the decision process is not known at this point, and it remains to be determined under what conditions a poor choice of priors will result in worse decisions than not using Bayesian methods at all. Clearly, further research is necessary to develop efficient methods for accurately assessing prior beliefs.

Using any one of a variety of input devices (i.e., optical scanning sheets, mark sense cards or computer cards) the examinees' test item responses are read by the computer and the Bayesian decision-theoretic procedure implemented. The computer program can be designed to provide the output necessary to monitor student progress through the instructional program. At a minimum, a statement of mastery allocations on objectives for each student can be produced, and this information can be used to guide a student through the next segment of his instruction.

The decision-theoretic procedure outlined in this paper provides a framework within which Bayesian statistical methods can be employed with criterion-referenced tests to improve the quality of decision-making in objectives-based instructional programs. The incorporation of losses introduces the decision-maker's values into the decision process. The Bayesian methods incorporate the prior knowledge of the decision-maker and utilize the data from all examinees, thereby effectively increasing the amount of information the decision-maker has without requiring the administration of additional test items. However, it should be pointed out that research is needed to establish the robustness of the Bayesian statistical model with respect to deviations of the data from the underlying assumptions. We also note that the Bayesian statistical model described in this paper is only one of several models that could be used (for example, see, Novick & Lewis, 1974, for another) within our decision-theoretic framework. Further study of these additional models would seem to be highly appropriate.

REFERENCES

BAKER, F. B. Computer-based instructional management systems: A first look. *Review of Educational Research*, 1971, **41**, 51-70.

GLASER, R., & NITKO, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement*. Washington: American Council on Education, 1971.

HAMBLETON, R. K. Testing and decision-making procedures for selected individualized instructional programs. *Review of Educational Research*, 1974, **44**, 371-400.

HAMBLETON, R. K., & NOVICK, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, **10**, 159–170.

LEWIS, C., WANG, M. M., & NOVICK, M. R. Marginal distributions for the estimation of proportions in *m* groups. *ACT Technical Bulletin No. 13*. Iowa City, Iowa: The American College Testing Program, 1973.

MILLMAN, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, **43**, 205–216.

MILLMAN, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications*. Berkeley, California: McCutchan Publishing Co., 1974.

NOVICK, M. R., & JACKSON, P. H. *Statistical methods for educational and psychological research*. New York: McGraw-Hill, 1974.

NOVICK, M. R., & LEWIS, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement*. Monograph Series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

NOVICK, M. R., LEWIS, C., & JACKSON, P. H. The estimation of proportions in *m* groups. *Psychometrika*, 1973, **38**, 19–45.

WANG, M. M. Tables of constants for the posterior marginal estimates of proportions in *m* groups. *ACT Technical Bulletin No. 14*. Iowa City, Iowa: The American College Testing Program, 1973.

AUTHORS

SWAMINATHAN, HARIHARAN. *Address:* School of Education, University of Massachusetts, Amherst, MA 01002. *Title:* Assistant Professor of Education and Psychology; Associate Director, Laboratory of Psychometric and Evaluative Research. *Degrees:* B. S. Dalhousie, M.Ed., M.S., Ph.D. University of Toronto. *Specialization:* Psychometric theory, multivariate statistics.

HAMBLETON, Ronald K. *Address:* School of Education, University of Massachusetts, Amherst, MA 01002. *Title:* Associate Professor of Education and Psychology; Director, Laboratory of Psychometric and Evaluative Research. *Degrees:* B.A. Waterloo, M.A., Ph.D. University of Toronto. *Specialization:* Psychometric theory, evaluation methodology.

ALGINA, JAMES J. *Address:* School of Education, University of Massachusetts, Amherst, MA 01002. *Title:* Research Associate, Laboratory of Psychometric and Evaluative Research. *Degrees:* B.A. University of Rhode Island, Ed.D. University of Massachusetts. *Specialization:* Psychometric theory, statistics.

*8.5 Simulation Study Involving Criterion-Referenced Test Scores

In this section, we present a paper by Hambleton, Hutten, and Swaminathan (1976). The authors compared four methods for estimating student mastery; these methods were discussed in section 8.3.2.

A COMPARISON OF SEVERAL METHODS FOR ASSESSING STUDENT MASTERY IN OBJECTIVES-BASED INSTUCTIONAL PROGRAMS[1]

RONALD K. HAMBLETON
LEAH R. HUTTEN
HARI SWAMINATHAN
University of Massachusetts

ABSTRACT

In objectives-based instructional programs where relatively short criterion-referenced tests are administered to estimate student mastery for the purpose of monitoring a student through the program, estimates which maximally utilize the information that can be obtained from the student during the alloted testing time are required. Bayesian estimates, which utilize prior information about the student, direct information provided by the student, and collateral information in the test data of other students, appear to be ideally suited for this purpose. In this paper, the relative merits of several methods, Bayesian and classical, for the estimation of student mastery are investigated. The effects of such factors as group homogeneity, test length, sample size, and prior information on the accuracy of the estimates as well as on decision-making accuracy are studied through computer simulations. it is shown that certain Bayesian estimates are superior to others, and the implications of the findings for objectives-based instructional programs are discussed.

ONE OF THE IMPORTANT PROBLEMS in objectives-based instructional programs such as Individualized Instruction (2, 3) concerns the assessment of student mastery. In order to monitor a student through an objectives-based program, pre-testing and post-testing is done to determine mastery on the specific program objectives (4). In theory at least, these tests can usually be made as reliable and valid as desired by increasing the test length. However, in practice, the total amount of testing time is limited and falls far short of the testing time needed to guarantee a high level of decision-making accuracy on the basis of test results. Needed are procedures that maximally utilize information that can be obtained from a student during the alloted time for testing. At present, there exist at least three approaches for doing this: tailored testing (1, 13); assessment of partial knowledge using new test scoring and/or test administration procedures (16); and assessment of student mastery using Bayesian statistical procedures (5, 7, 11, 12, 14).

The possibility of using Bayesian statistical procedures for the assessment of student mastery is particularly appealing because they require absolutely no change from the usual test administration procedures. Moreover, a careful examination of the Bayesian theory suggests that some meaningful gains may be realized using Bayesian methods (11). Finally, empirical work with similar Bayesian methods in another application confirmed the efficacy of this class of Bayesian methods (10).

In the typical objectives-based program, relatively short criterion-referenced tests are used to determine student mastery. A cutoff score or threshold score is set to permit the decision-maker to assign examinees, on the basis of their performance on each subset of items measuring an objective covered in the criterion-referenced test, into one of two mutually exclusive categories—masters and non-masters. Usually the proportion-correct score for an examinee on items measuring each objective is compared to the cutting score for the purpose of decision-making. However, the method of using proportion-correct scores as estimates of mastery is not entirely satisfactory when the number of items on which the proportions is based is few and when there are many students. In situations where one is interested in estimating many mastery scores, some, by chance, will be substantially overestimated and others underestimated. The implication is that many of the decisions made on the basis of test results will be incorrect, and this reduces the overall effectiveness of the instructional program.

The Bayesian procedure, in theory at least, utilizes additional information on each student that is available but ignored by non-Bayesian procedures. According to Novick (9), this is done by using not only the *direct information* provided by a student's test score but also using the *collateral information* contained in the test data of other students and any *prior information* on the student that may be available.

In view of the current interest in applying Bayesian methods to testing problems within the context of objectives-based programs, the purpose of the present investigation was to study, in a systematic way, the relative merits of several methods for estimating student mastery. In addition to the proportion-correct score, the classical Model II estimate given by Jackson (6), and Bayesian estimates such as the Bayesian joint modal estimate (12), the marginal mean estimate (7) and a modification of it were studied.

The "quality" of the various estimates are dependent, to some extent, on the interactive effects of factors such as group homogeneity, test length, sample size, and, for the Bayesian estimates, the prior information available on the ability of the examinees. For this reason, a computer simulation study was conducted where we produced comparisons of "true" mastery scores with "estimated" mastery scores obtained by the methods described above for different ability distributions, test lengths, sample sizes, and for varying amounts of prior information on the examinees.

Method

Four Methods of Estimating Student Mastery

Proportion-Correct Estimate

The simplest and the most obvious estimate of the ith examinee's true mastery score, π_i, defined as the proportion of items in the domain of items measuring the objective that the examinee can answer correctly, is his observed proportion score, $\hat{\pi}_i$. This estimate is obtained by dividing the examinee's test score, x_i (the number of items answered correctly), by the total number, n, of the items measuring the objective included in the test. Appealing as it may seem in view of the fact that the proportion-correct score is an unbiased estimate of the true mastery score, this estimate, as mentioned earlier, is extremely unreliable when the number of items on which the estimate is based is small. For this reason, procedures that take into account other available information in order to produce improved estimates, especially in the case when there are few items measuring an objective in the test, would be more desirable.

Classical Model II Estimate (Jackson)

One of the first attempts to produce an estimate of the true score of an examinee using the information obtained from the group to which an individual belongs was made by Kelley in 1927. This is the well-known regression estimate of true score (8:65), which is the weighted sum of two components—one based on the examinee's observed score and the other based on the mean of the group to which the examinee belongs. Jackson (6) modified this procedure for use with binary data by transforming the test score x_i and g_i via the arcsine transformation, known as the Freeman-Tukey transformation,

$$g_i = \frac{1}{2}\left(\sin^{-1}\sqrt{\frac{x_i}{n+1}} + \sin^{-1}\sqrt{\frac{x_i+1}{n+1}}\right) . \qquad (1)$$

As a result of this transformation, the true mastery score is transformed into γ_i, where,

$$\gamma_i = \sin^{-1}\sqrt{\pi_i} . \qquad (2)$$

If π_i is not too large or too small, and if n, the number of test items, is sufficiently large, then the distribution of g_i is approximately normal with a mean approximately equal to the transformed true mastery score, γ_i, and known variance

$$\phi = (4n+2)^{-1} .$$

The Model II estimate, or the modified Kelley estimate, becomes, in terms of γ,

$$\hat{\gamma}_i = [g_i\hat{\phi}_\gamma + (4n+2)^{-1}g_{\cdot}] / [\hat{\phi}_\gamma + (4n+2)^{-1}] \qquad (3)$$

where $g_{\cdot}$, the sample mean based on a sample of N examinees, is given by

$$g_{\cdot} = N^{-1}\sum_{i=1}^{N} g_i , \qquad (4)$$

and $\hat{\phi}_\gamma$, the sample variance of the γ's, is given by

$$\hat{\phi}_\gamma = (N-1)^{-1}\sum_{i=1}^{N}(g_i - g_{\cdot})^2 - (4n+2)^{-1} . \qquad (5)$$

Once $\hat{\gamma}_i$ is obtained, $\hat{\pi}_i$ is determined from the expression

$$\hat{\pi}_i = (1 + .5/n)\sin^2\hat{\gamma}_i - .25/n . \qquad (6)$$

For a detailed discussion of this estimate, the reader is referred to Novick, Lewis, and Jackson (12).

Bayesian Joint Modal Estimate

The Jackson estimate given above is not the ideal estimate since it does not take into account any prior information that may be available. In addition, it may happen that $\hat{\phi}_\gamma$ estimated using Equation 5 is negative, in which case the solution will not be meaningful. Novick et al. (12), utilizing the transformation 1 and 2, obtained a Bayesian solution for the estimation of the mastery score that not only takes into account the direct and collateral information but also any prior information that may be available. In addition, this procedure avoids the problem of negative estimates for ϕ_γ.

BEST COPY AVAILABLE

The Bayesian solution is more complicated that the classical Model II solution and involves an iterative procedure. The Bayesian procedure assumes that, in addition to the assumption that g_i is normally distributed with mean γ_i and variance $(4n + 2)^{-1}$, the transformed true mastery scores $\gamma_1, \gamma_2, \ldots, \gamma_N$ of the N individuals come from a normal population with unknown mean a and unknown variance ϕ_γ. Thus, to use the Bayesian procedure, the prior knowledge about a and ϕ_γ have to be specified.

However, Novick et al. (12) suggest that prior knowledge about a may not be as important as the specification of prior beliefs about ϕ_γ. Furthermore, they have suggested that it is reasonable to represent the prior belief about ϕ_γ by an inverse chi-square distribution, which depends on only two parameters, υ and λ (12). Thus, in order to indicate one's prior belief one has to specify υ and λ (12). (For details of this procedure, see 12, 14.)

Marginal Mean Estimate

The Bayesian Model II estimate discussed above is in reality the joint modal estimate. This joint estimate is useful for making joint decisions about a set of N examinees. However, in criterion-referenced testing situations, separate decisions about each individual have to be made, and, hence, separate or marginal estimates of true mastery scores are required.

Lewis, Wang, and Novick (7) have obtained a marginal mean estimate of the true mastery score, given by

$$\hat{\gamma}_i = g_{\cdot} + \rho^*(g_i - g_{\cdot})\,. \tag{7}$$

The quantity ρ^* is dependent on the parameters υ and λ and on the data; once the parameters are set, ρ^* can be read directly from tables prepared by Wang (15). Again, once $\hat{\gamma}_i$ is obtained, $\hat{\pi}_i$ is determined using Equation 6.

In obtaining the joint modal estimates and the marginal mean estimates, the above authors assumed that the prior beliefs about a and ϕ_γ could be expressed in the form of distributions. In the present investigation, it was felt that the effects on the marginal mean estimates of specifying the prior beliefs about a and ϕ_γ as point values should also be studied. To this end, we obtained marginal mean estimates based on the assumptions that (*1*) the prior belief about a can be expressed as a uniform distribution, but that ϕ_γ can be specified exactly; and (2) both a and ϕ_γ can be specified exactly. In the first case, it can be shown that the marginal mean estimate $\hat{\gamma}_i$ is given by

$$\hat{\gamma}_i = \frac{g_i\phi_\gamma + (4n + 2)^{-1}\, g_{\cdot}}{\phi_\gamma + (4n + 2)^{-1}} \tag{8}$$

In the second case, the marginal mean estimate, $\hat{\gamma}_i$, becomes

$$\hat{\gamma}_i = \frac{g_i\phi_\gamma + (4n + 2)^{-1}\, a}{\phi_\gamma + (4n + 2)^{-1}} \tag{9}$$

The similarity between the marginal mean estimates given by Equations 8 and 9 and the Jackson estimate given by Equation 3 is obvious.

Factors Under Consideration

Sample Size

Samples of examinees of size 15, 25, and 50 were considered. These values were selected because they seemed to be typical of the class sizes that might be expected to occur in practice. (Selected results with sample sizes larger than 50 were obtained, but they were essentially the same results obtained with smaller sample sizes.)

Test Length

Tests of length 8, 10, and 20 items were employed in the simulations. Novick et al. (12) recommend the use of their methods when the test includes at least eight items. Twenty items represents a reasonable upper limit on the number of items to be used to measure the mastery of an objective covered in a criterion-referenced test.

Homogeneity of the True Mastery Score Distribution

Two fairly typical distributions of true mastery scores were considered. To obtain a homogeneous distribution it was assumed that the true mastery scores were distributed in the following way: 20% of the population were distributed uniformly on each of the five intervals defined by the end points .46 and 1.00 and the four middle boundary points .70, .79, .85, and .91. These values were selected to correspond roughly to a beta distribution with mean .80 and variance .145. The beta distribution rather than the normal distribution was chosen because the beta distribution is defined on the same interval [0, 1] as the true mastery scores, whereas the normal distribution is not.

To obtain a heterogeneous distribution, we assumed that 20% of the population were distributed uniformly in each of the five intervals defined by the end points .30 and 1.00 and the four middle boundary points, .60, .70, .80, and .90.

Specification of Prior Information

An integral part of the process of utilizing Bayesian methods is the specification of one's prior belief about the distribution of true mastery scores. Unfortunately, the importance of the specification of a prior under varying testing conditions, for example, tests of different lengths administered to different numbers of examinees, is unknown. In our study, since the distribution of true mastery scores in the population was known, it was possible to investigate the effects of specifying prior information on the various Bayesian estimates. Since the specification of prior information depends upon setting the value of ϕ_γ, we studied the effects of setting four different values of ϕ_γ on the Bayesian joint modal and Bayesian marginal mean estimates. The four situations were: an accurate value of ϕ_γ based on the distribution of transformed true mastery scores; a large value for ϕ_γ; a small value for ϕ_γ; and a value for ϕ_γ derived from the data.

In addition, with the modified mean estimate, we generated four variations on the setting of prior information. We set ϕ_γ to be one of two values: an accurate value of ϕ_γ based on the distribution of transformed true mastery scores and a small value for ϕ_γ. Also, a was set to be one of two values: a value based on the examinee's simulated performance with true mastery score π on one previous test or on five previous tests.

Cutting Score

In this study the cutting score was set to be .80. This value of the cutting score is often observed in practice.

Testing the Fit of the Data

Two criteria were used to test the goodness of fit between the various estimates and the true mastery scores. These criteria are the loss functions based on (*a*) the average absolute difference and (*b*) decision accuracy. These two were selected because they seemed to be the most relevant within the context of criterion-referenced testing problems.

Average Absolute Difference (AAD)

This loss function is based on the average absolute difference between the estimates and the true mastery scores, and is given by the expression

$$\text{AAD} = \sum_{i=1}^{N} |\hat{\pi}_i - \pi_i| / N .$$

Decision Accuracy

The second measure of goodness of fit used in our study was the proportion of correct and incorrect decisions arrived at using the various estimates of mastery scores.

When an examinee's true mastery score π_i exceeds π_o, where π_o is the point on the mastery score scale used to separate examinees into mastery and non-mastery states, the examinee is considered a true master. Likewise, when the true mastery score is below the cutoff score, the examinee is considered a true non-master. Since in practice these true mastery scores are not known, the allocation of examinees to mastery states is based upon observed scores, or the estimates π_i of the true mastery scores π_i. Again, if the estimated true mastery score π_i exceeds π_o, the examinee is classified as a master, and a non-master otherwise.

In order to investigate the goodness of fit for the various estimates of the true mastery score and to study how these estimates affect the classification of examinees, we define a variable Y_i such that

$$Y_i = 1, \quad \text{if} \quad \pi_i \geqslant \pi_o$$

and

$$Y_i = 0, \quad \text{if} \quad \pi_i < \pi_o .$$

If we obtain an estimate $\hat{\pi}_i$ of π_i, then we can define the "estimate" of Y_i as

$$\hat{Y}_i = 1, \quad \text{if} \quad \hat{\pi}_i \geqslant \pi_o$$

and

$$\hat{Y}_i = 0, \quad \text{if} \quad \hat{\pi}_i < \pi_o .$$

The error of estimation can then be defined as $e_i = (\hat{Y}_i - Y_i)$. Obviously, the error e_i can take on one of three values, -1, 0, +1. When an examinee is classified correctly, $e_i = 0$. When a false positive error is commited, that is, when an examinee who is a true non-master is classified as a master, $e_i = 1$. Similarly, if a false negative error is committed, that is, when a true master is classified as a non-master, $e_i = -1$. We then define decision accuracy as

$$1 - \sum_{i=1}^{N} e_i^2 / N ,$$

the ratio of the number of correct decisions to the total number of correct decisions to the total number of decisions made.

In passing, it should be pointed out that the false positive and false negative errors can be weighted differentially, or in other words, various costs of misclassification can be introduced (5). In our study, losses were weighted equally, although it should be recognized that in many applications different losses will be used.

Simulating the Test Data

The first step in the simulation of test data was to specify the number of examinees, the test length, and the true mastery score distribution. The next step was to generate a true mastery score for each examinee. This was accomplished by selecting a number at random from the true mastery score distribution.

The third step was to generate a sample mastery or proportion-correct score for each examinee. The simulation of test data was accomplished using the binominal test model (8:508). Since the true mastery score π for

Table 1.–Goodness of Fit Results[1] Based on the Average Absolute Deviation Measure for 25 Simulations of Each of Three Sample Sizes in which Homogeneity of the Ability Distribution, Number of Test Items, and Prior Information Were Varied

Estimate	Ability Distribution					
	Heterogeneous			Homogeneous		
	Test Length			Test Length		
	8	10	20	8	10	20
Proportion-Correct	.112	.098	.075	.103	.093	.067
Jackson	.092	.085	.069	.080	.075	.056
Modified Marginal						
Mean -1	.085	.077	.057	.082	.073	.050
-2	.078	.067	.054	.060	.053	.047
-3	.094	.083	.051	.097	.087	.050
-4	.056	.047	.037	.049	.043	.032
Bayesian Joint						
Modal -1	.111	.100	.072	.093	.087	.062
-2	.104	.098	.066	.089	.085	.061
-3	.133	.127	.074	.103	.103	.087
-4	.092	.081	.063	.074	.069	.056
Bayesian Marginal						
Mean -1	.093	.086	.070	.081	.076	.057
-2	.090	.084	.064	.076	.073	.057
-3	.099	.089	.066	.085	.081	.060
-4	.093	.081	.063	.076	.071	.056

---

[1] Smaller values indicate better estimates.

the examinee represented the probability of correctly answering any item, it was possible to convert the probability into item scores (i. e., 1 for a correct response and 0 for an incorrect response) by comparing π with random numbers selected from a uniform distribution on the interval $[0, 1]$. If the random number was less than or equal to π, the examinee was credited with a correct response; otherwise, the examinee was credited with an incorrect response. This process was repeated n times (for n items) and a test score for the examinee was obtained. This test score was then converted to a proportion-correct score, and the procedure was repeated for each of the N examinees.

Once the proportion-correct scores for the sample of N examinees on n items were obtained, "improved" estimates were obtained by the methods described earlier. Each set of estimates derived from the different methods for the three test lengths, three sample sizes, and from the two ability distributions was then compared with the "true" values. Goodness of fit measures described earlier were used to assess the appropriateness of each set of estimates. To improve the stability of the goodness of fit measures, 25 replications were conducted for each set of test conditions. We reported the mean goodness of fit measures across the 25 replications.

Results

In Tables 1 and 2, we have reported the results of our simulations. The entries in the tables require some explanation. The modified marginal mean estimates 1, 2, 3, and 4 were obtained using Equation 9. These four estimates differ with respect to the specification of a and ϕ_γ. In estimates 1 and 3, a was set equal to $\sin^{-1}\sqrt{\hat{\pi}_i}$ where $\hat{\pi}_i$ was an estimate of the ith examinee's true mastery score derived from simulating his performance on one previous test occasion. For estimates 2 and 4, the same procedure was followed with one exception: $\hat{\pi}_i$ was the examinee's average test performance on five previous tests. With estimates 1 and 2, ϕ_γ was set equal to the variance of γ, obtained from the true distribution of γ. For estimates 3 and 4, the same procedure was followed with one exception: ϕ_γ was set to be ¼ of the variance of γ.

Since the effects of prior information on the Bayesian joint modal and the marginal mean estimates were to be investigated, we chose to consider four different priors. The prior information for the two Bayesian estimation procedures is summarized by specifying the two parameters, υ and λ. Novick, Lewis, and Jackson (12) recommend that it is appropriate in most situations to set $\upsilon = 8$ and $\lambda = 6\phi_\gamma$. Hence, varying the prior information

BEST COPY AVAILABLE

amounts to setting different values for ϕ_γ. In our study, the Bayesian joint modal and the Bayesian marginal mean estimates 1, 2, 3, and 4 were obtained by setting ϕ_γ equal to one of four values. For the first estimate, ϕ_γ was set equal to the variance of the transformed observed proportion scores in the sample of examinees. With the second estimate, ϕ_γ was taken to be the variance of γ_i derived from Equation 2. In other words, ϕ_γ was set equal to the variance of the transformed true mastery scores. For the third estimate, ϕ_γ was set to be ¼ times as large as the variance of the transformed true mastery scores. For the fourth estimate, ϕ_γ was set to be 4 times the variance of the transformed true mastery scores. These last two estimates were introduced so that we could study the effects of a prior based on a distribution that was either too homogeneous or too heterogeneous relative to the distribution of mastery scores in the sample of examinees.

Discussion

Sample Size

Since the results of our simulation study varied only slightly as a function of sample size, we chose to simplify the presentation of results by reporting average goodness of fit measures across the three sample sizes for the remaining six combinations of test lengths and true mastery score distributions.

Test Length

Increasing the test length improved both the goodness of fit measures; however, the improvements were modest. On the average, and across all of the estimates, there was an 8% improvement in decision-accuracy and a decrease of about .027 in the average absolute difference index when the test length was increased from 8 to 20 items. This suggests that 8 items represents a sufficient basis on which to assess student mastery or to make instructional decisions from criterion-referenced test data.

Group Homogeneity

The group homogeneity had some interesting effects on the goodness of fit measures. The goodness of fit measure based on the average absolute deviations indicated that the estimation procedures were more effective with the homogeneous group than with the heterogeneous group. This perhaps can be explained when one realizes that the more homogeneous the distribution of ability, the more valuable the group mean is for the estimation of an individual's mastery score.

However, in terms of decision-making accuracy, the situation was reversed. The decision-making accuracy was better for a heterogeneous ability distribution than for a homogeneous distribution. An explanation for this is as follows: In a homogeneous distribution where the true mastery scores are concentrated near the mean, and in

Table 2.–Goodness of Fit Results[1] Based on a Decision Accuracy Measure for 25 Simulations of Each of Three Sample Sizes in which Homogeneity of the Ability Distribution, Number of Test Items, and Prior Information Were Varied

Estimate	Ability Distribution					
	Heterogeneous			Homogeneous		
	Test Length			Test Length		
	8	10	20	8	10	20
Proportion-Correct	.827	.801	.858	.763	.776	.819
Jackson	.807	.824	.860	.737	.754	.794
Modified Marginal						
Mean -1	.831	.847	.884	.815	.822	.831
-2	.838	.865	.887	.853	.840	.868
-3	.821	.837	.903	.794	.807	.339
-4	.871	.905	.922	.864	.858	.908
Bayesian Joint						
Modal -1	.730	.782	.853	.656	.696	.800
-2	.792	.782	.849	.700	.738	.802
-3	.654	.642	.843	.622	.625	.691
-4	.793	.829	.859	.776	.744	.808
Bayesian Marginal						
Mean -1	.801	.821	.858	.725	.758	.802
-2	.808	.811	.849	.738	.761	.806
-3	.765	.806	.856	.744	.734	.788
-4	.793	.829	.859	.780	.737	.809

---

[1] Larger values indicate better estimates.

BEST COPY AVAILABLE

our study near the cutting score, more incorrect decisions are bound to occur. This is because all the estimates, with the exception of the proportion correct scores, weight the observed score by the group mean, and even the slightest change in either direction of the cutting score would make the estimate unstable in terms of decision-making accuracy. This occurs to a lesser extent in heterogeneous distributions because of the spread of scores that exists.

Prior Information

Specification of prior information was required only for the Bayesian estimates. We considered the effects of varying priors on all three Bayesian estimates; namely, the modified marginal mean estimate, the joint modal estimate, and the marginal mean estimate. Since setting the prior for the modified marginal mean estimate required a procedure that was different from that required by the joint modal and the marginal mean estimates, we shall consider them separately.

An examination of the entries in Table 1 indicates that for the Bayesian joint modal estimate and the marginal mean estimate, in general, estimate 4 produced the best results, followed by estimates 2, 1, and 3. For estimate 4, the prior information was based on a value of the variance that was four times the true variance, while for estimate 3 the prior information was based on a value of the variance that was (.25) times the true variance. Since estimate 1 was based on the sample variance, we shall not be concerned with it for the present. Estimate 2 was based on the variance of the true mastery score distribution; in our case, it was taken to be .04.

The expression for the marginal mean estimate is given by Equation 7. The quantity ρ^* in Equation 7 increases with λ/υ, and when ρ^* is large, little weight is given to the group mean and vice versa. Thus, when ϕ_γ was set equal to .16 (= 4 × .04), the value of ρ^* is large and there is little or no regression towards the mean of the group. Hence, the estimate for γ_i is close to the mean, or the true score, of that examinee. Since our criterion is based on the deviation of the estimated mastery score from the true score, a smaller value for error would be obtained in this case. A small value for λ/υ—as that obtained when for estimate 3, ϕ_γ was set equal to .01 (= .25 × .04)—results in considerable regression towards the mean of the group. If an examinee's true score is far from the mean of the group, a larger value for error would be obtained. This fact is borne out by the fact that in the relatively homogeneous distribution, the errors are relatively small. A similar explanation is valid for the joint modal estimates. In general, prior information had little effect on the estimates as test length increased.

The modified marginal mean 4, on the other hand, produced best results when ϕ_γ was set equal to .25 of the variance of the true distribution of γ and when α was set equal to the examinee's average score on five previous tests. The reason for this is obvious when we examine Equation 9. When ϕ_γ is small, i. e., when the reliability of the test is low, the estimate weights α rather heavily. In this case, an α based on the examinee's five previous test scores is an extremely good initial estimate of the examinee's true score, and hence we obtained excellent results with this estimate.

In summarizing, we note that in the present context, the Bayesian joint modal and marginal mean estimates are less affected by the prior information than the modified marginal mean estimates unless a really bad prior is specified. If exact and accurate values for α and ϕ_γ can be specified, then the modified marginal mean estimates produce the best results. Since the modified marginal mean estimates are sensitive to priors, care should be exercised in using them.

Comparison of the Estimates

The results indicated that all the estimates fared far better than the proportion-correct score in terms of the average absolute deviation measure. The best estimates were obviously the modified marginal mean estimates. The other estimates, in order, are: (1) the marginal mean and Jackson estimates; (2) the joint modal estimate based on a good prior; (3) the proportion-correct score; and (4) the joint modal estimate based on a bad prior.

In terms of decision-making accuracy, the results were less clear cut. The modified mean estimates were again the best. The proportion-correct score, the Jackson, and the marginal mean estimates followed, in that order. The joint modal estimates, though not too far behind, produced the poorest results.

Possible explanations for the poor results obtained with the joint modal estimate were given in the previous sections. The explanation is that the joint modal estimates are strictly intended for making joint decisions about all the examinees. However, our criteria, the average absolute deviation measure and the decision-accuracy measure, are both based on the deviation of individuals from their true mastery scores, and hence are biased against the joint modal estimate.

Summary and Suggestions for Further Research

In summary, the results of our simulation study comparing several methods for estimating student mastery in a variety of testing situations were rather revealing. Specifically, the classical Model II estimate and the Bayesian estimates tended to produce better results than the proportion-correct estimate, and we obtained better results with the Bayesian estimates when the distribution of true mastery scores was homogeneous. Also, we noted that test length, and amount of prior information, had only minor effects on the "quality" of the estimates. The one exception occurred with the Bayesian modal estimate for which the prior, set on the true mastery score distribution, was too homogeneous. In this situation the results were quite poor.

In comparing the estimates, we noted that the modified marginal mean estimates tended to be "best," but this result is somewhat misleading since in practice one would seldom have as good a prior estimate of the examinee's level of mastery and the distribution of true mastery scores as we used in the study.

Although the results of our simulation study revealed only modest improvements with Bayesian methods under the conditions studied, we are not prepared to discourage the use of Bayesian methods with criterion-referenced test data. Quite the contrary, since with the availability of a small computer, any improvement in the estimates, however modest, is worth obtaining, especially as these can be obtained with very little cost and effort.

It should be mentioned also that Bayesian methods can be used to produce a posterior distribution on the unknown true mastery score for an examinee which, when incorporated with a loss structure, provides a basis for decision-making (7, 14). Our results did indicate that, on the average, better point estimates are obtained from the Bayesian methods. Hence, Bayesian procedures do provide a better basis for generating a probability distribution to represent our belief about the location of the unknown true mastery score for the individual.

Also, we believe that as users become more adept at stating prior beliefs about examinees' level of mastery, and if the Bayesian modal estimate is avoided in situations when individual descriptions or decisions are required, even better results than those reported in this paper will be obtained. We should add that we would expect the decision-accuracy associated with Bayesian methods to improve in situations where there are more than two mastery states (14).

In terms of further research, we think it highly desirable to explore the possibility of applying the Bayesian methods to tests with fewer than eight items. This would make the methods applicable to many more testing situations than is possible now. Also, it is with the shorter tests that improvements on the proportion-correct estimates are most needed.

NOTE

1. Without in any way implying their endorsement of the research methodology used in the study or the results, the authors would like to acknowledge their gratitude to James Algina, Paul Jackson, and Melvin Novick for constructive criticisms and helpful comments on an earlier draft of the manuscript. Ming-Mei Wang kindly provided us with a computer program to compute the marginal mean estimates.

REFERENCES

1. Ferguson, R. L. The development, implementation, and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.
2. Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist*, 1963, *18*, 519-521.
3. Glaser, R. Adapting the elementary school curriculum to individual performance. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1968.
4. Hambleton, R. K. Testing and decision-making procedures for selected individualized instructional programs. *Review of Educational Research*, 1974, *44*, 371-400.
5. Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159-170.
6. Jackson, P. H. Simple approximations in the estimation of many parameters. *British Journal of Mathematical and Statistical Psychology*, 1972, *25*, 213-229.
7. Lewis, C., Wang, M. M., & Novick, M. R. Marginal distributions for the estimation of proportions in *m* groups. *Psychometrika*, 1975, *40*, 63-75.
8. Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.
9. Novick, M. R. Bayesian considerations in educational information systems. In *Proceedings of the 1970 Invitational Conference on Testing Problems.* Princeton, N. J.: Educational Testing Service, 1971.
10. Novick, M. R., Jackson, P. H., Thayer, D. T., & Cole, N. S. Applications of Bayesian methods to the prediction of educational performance. *The British Journal of Mathematical and Statistical Psychology*, 1972, *25*, 33-50.
11. Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: Center for the Study of Evaluation, University of California, 1974.
12. Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in *m* groups. *Psychometrika*, 1973, *38*, 19-46.
13. Spineti, J., & Hambleton, R. K. A computer simulation study of tailored testing strategies for objectives-based instructional programs. *Educational and Psychological Measurement*, in press.
14. Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement*, 1975, *12*, 87-98.
15. Wang, M. M. Tables of constants for the posteriori marginal estimates of proportions in *m* groups. *ACT Technical Bulletin No. 14.* Iowa City, Ia.: The American College Testing Program, 1974.
16. Wang, M. D., & Stanley, J. C. Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 1970, *40*, 663-705.

8.6 Reporting of Test Score Information

In this section, we will discuss some procedures for reporting individual and group data from criterion-referenced tests. The examples included in each sub-section are based upon the involvement of one of the authors with criterion-referenced testing programs in several school systems. The examples provide a look at a number of ways of reporting individual and group test score information.

For the reader interested in a further discussion of how to report test score information, the books by Gronlund (1974, 1976) provide an excellent review of practical procedures.

8.6.1 Individual Test Score Information

First, we will provide some examples of how to report individual test score information, and then we will discuss an alternative method of presentation that does not involve the usually reported percentage correct by objectives.

The examples that follow present test results of three students on three tests. The data presented is the percentage correct score for each objective on each of the three tests, where the tests have a varying number of objectives. Also, data is presented on the average performance across the objectives and the percent of the objectives mastered across three test occasions. Two comments follow from a perusal of the data: (1) This data output for an individual student provides an excellent breakdown of performance, and is highly useful for decision-making on an individual basis. (2) A reasonable way to present individual data is by using a percentage correct score. In reference to comment two, we will now discuss an alternate method for presenting individual test score data.

Green Valley--Title I   SCHOOL = Green Valley   GRADE = 5   DATE = June 1977

```
**************************************************************************************************
STUDENT ID = 1 16  5 09535  0  STUDENT NAME =  9535              DATE =    JUNE 1977        FOLLOWUP TESTING = N/A
TOTAL NUMBER OF TESTS ADMINISTERED =3                             BACKGROUND DATA = STSTA  0 SEX    2 ETHNO  3        4

                                       SUMMARY OF PERCENTAGE SCORES (PS) BY OBJECTIVE                     NUMBER OF      MAXIMUM
  TESTS/OBJECTIVES          1    2    3    4    5    6    7    8    9   10   11   12   13   14   15  OBJECTIVES       SCORE

WORD ATTACK LEVEL 3        67   67  100   95   89  100  100   75  100   89                               10             107
WORD ATTACK LEVEL 4       100  100  100   75   89   67   88  100   75  100  100   88  100                13             105
DICTIONARY                 20  100   80   11   40                                                        5              38

TEST OCCASION          AVERAGE PERFORMANCE ON OBJECTIVES                           PERCENT OF OBJECTIVES MASTERED
                 P-RD  WA-1  WA-1M  WA-2  WA-3  WA-4  DICT  R-CMP            P-RD  WA-1  WA-1M  WA-2  WA-3  WA-4  DICT  R-CMP

  FIRST                                    65    77    65                                                10    46    60

  SECOND                                   89                                                            80

  THIRD                                    90    91    50                                                90    77    40

**************************************************************************************************
**************************************************************************************************
STUDENT ID = 1 16  5 09538  0  STUDENT NAME =  9538              DATE =    JUNE 1977        FOLLOWUP TESTING = N/A
TOTAL NUMBER OF TESTS ADMINISTERED =3                             BACKGROUND DATA = STSTA  0 SEX    2 ETHNO  3        3

                                       SUMMARY OF PERCENTAGE SCORES (PS) BY OBJECTIVE                     NUMBER OF      MAXIMUM
  TESTS/OBJECTIVES          1    2    3    4    5    6    7    8    9   10   11   12   13   14   15  OBJECTIVES       SCORE

WORD ATTACK LEVEL 3        67   67  100   75   89   83  100  100   86   67                               10             107
WORD ATTACK LEVEL 4       100  100   56   63   89   67  100   83   89  100  100  100   83                13             105
DICTIONARY                100  100  100   39   60                                                        5              38

TEST OCCASION          AVERAGE PERFORMANCE ON OBJECTIVES                           PERCENT OF OBJECTIVES MASTERED
                 P-RD  WA-1  WA-1M  WA-2  WA-3  WA-4  DICT  R-CMP            P-RD  WA-1  WA-1M  WA-2  WA-3  WA-4  DICT  R-CMP

  FIRST                                    75    83    51                                                40    54    40

  SECOND                                   87                                                            80

  THIRD                                    88    87    80                                                80    77    60

**************************************************************************************************
**************************************************************************************************
STUDENT ID = 1 16  5 09615  0  STUDENT NAME =  9615              DATE =    JUNE 1977        FOLLOWUP TESTING = N/A
TOTAL NUMBER OF TESTS ADMINISTERED =3                             BACKGROUND DATA = STSTA  0 SEX    1 ETHNO  3        5

                                       SUMMARY OF PERCENTAGE SCORES (PS) BY OBJECTIVE                     NUMBER OF      MAXIMUM
  TESTS/OBJECTIVES          1    2    3    4    5    6    7    8    9   10   11   12   13   14   15  OBJECTIVES       SCORE

WORD ATTACK LEVEL 3        93   67  100   81   78  100  100   75  100   89                               10             107
WORD ATTACK LEVEL 4       100   63   67   75  100  100  100   83   89   78  100   63  100                13             105
DICTIONARY                100  100  100   56   40                                                        5              38

TEST OCCASION          AVERAGE PERFORMANCE ON OBJECTIVES                           PERCENT OF OBJECTIVES MASTERED
                 P-RD  WA-1  WA-1M  WA-2  WA-3  WA-4  DICT  R-CMP            P-RD  WA-1  WA-1M  WA-2  WA-3  WA-4  DICT  R-CMP

  FIRST                                    74    76    74                                                30    54    60

  SECOND                                   89                                                            70

  THIRD                                    88    86    79                                                70    62    60

**************************************************************************************************
```

BEST COPY AVAILABLE

In sections 8.3.2 and 8.4.2, we discussed the use of Bayesian procedures for estimating domain scores and for making decisions about assignment to mastery states. If Bayesian procedures are used for the above two purposes, Ferguson and Novick (1973) have discussed the practical feasibility of presenting individual data [illegible] a different form than simply percent correct. Rather than simply presenting percentage correct by objective, Ferguson and Novick call for a change to new procedures such that:

> Under the proposed changes, rather than evaluating student proficiency solely on the posttest results, additional data would be incorporated within the decision analysis process, and furthermore, the quantity reported would be an index relating the student's estimated proficiency to a stipulated standard. However, it should be emphasized that although the nature of the data reported in the student profile would change, the procedures employed by the teacher and/or student to judge proficiency would remain the same.

Thus, by employing Bayesian procedures, an alternate way of presenting individual data by objective can be utilized. This involves the use of a mastery index. Before discussing the index and its interpretation, some data, similar to the data in Ferguson and Novick (1973), will be presented.

Objective	Percent Correct	Mastery Index
1	87.5	80
2	75	76
3	100	92

In order to discuss the mastery index, a relevant cut-off point, using one of the suggested methods in Unit 6, must be established. Assume that it is .85. The mastery index for each objective then gives the probability that the student's level of proficiency is above .85. For instance, on objective 2, the student got 75% of the items correct, which is less than the cut-off of .85. However, when the collateral and prior information is combined with the percent correct information, we get a probabilistic statement: that although his/her percent correct score is below the cut-off, we are still 76% certain his/her domain score is above .85. For this student, it is apparent that his/her performance on the test is lower than the performance on the collateral data being used.

In sum, Bayesian analysis provides a probabilitic statement about mastery. The mastery index gives the test score interpreter a probability of success figure, while the percent correct has no probabilistic interpretation attached; it is either above or below the cut-off. How might this probability statement be used? Suppose, for instance, you were willing to move a student on to another objective if the odds were better than two to one in favor of his/her actually being proficient. Then the student would be advanced if his/her probability of mastery was greater than .67. For the student in our example, he/she would be passed to the next objective even though his/her percent correct score was less than the cut-off. Rather than making a yes-no decision, the mastery index allows you to ascertain the probability that you are making the correct decision. Of course, the correctness of the probability statement will depend on the quality of the collateral and prior information used in obtaining "revised" domain score estimates.

8.6.2 Group Test Score Information

In this section, two examples of methods for presenting group test data will be discussed. Then a helpful table for making decisions about group sizes will be presented and discussed.

The first example is actually a set of examples, based upon group data for the school system discussed in the last sub-section.

The first set of tables gives a district summary of performance on each reading objective for 6 tests. Average percent scores and percentage of examinees who mastered are presented for each objective for each test. The first table is collapsed across all 6 grades; subsequent tables give district information, but summarized by grade (two examples are given).

SYSTEM SUMMARY OF PERFORMANCE ON EACH READING OBJECTIVE REPORTED
FOR EACH TEST ADMINISTERED

DISTRICT SUMMARY OF TEST RESULTS DATE JUNE 1977

TESTS/OBJECTIVES	SUMMARY OF AVERAGE PERCENTAGE SCORES BY OBJECTIVES 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE	NUMBER OF EXAMINEES
PREREADING	92	87	95	97	87	95	95	97	91	81	86	89	71	70		88.6	255
WORD ATTACK LEVEL 1	96	92	91	84	75	90										88.3	491
WORD ATTACK LEVEL 2	93	89	89	80	86	72	88	66	87	69						81.3	504
WORD ATTACK LEVEL 3	74	73	78	52	84	79	87	62	69	67						77.5	631
WORD ATTACK LEVEL 4	78	83	71	67	82	79	78	85	63	73	84	75	73			75.7	425
DICTIONARY	88	87	77	47	49											69.7	358

TESTS/OBJECTIVES	PERCENTAGE OF EXAMINEES WHO MASTERED OBJECTIVES 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE
PREREADING	72	72	90	98	61	81	87	93	88	62	72	84	64	54		76.5
WORD ATTACK LEVEL 1	95	86	87	66	63	87										80.6
WORD ATTACK LEVEL 2	84	78	78	68	81	61	58	34	77	55						67.8
WORD ATTACK LEVEL 3	55	44	52	68	63	53	77	25	78	18						54.3
WORD ATTACK LEVEL 4	59	69	76	27	59	52	59	81	26	37	81	55	51			57.2
DICTIONARY	85	85	72	17	19											56.0

BEST COPY AVAILABLE

DISTRICT SUMMARY OF TEST RESULTS GRADE 1 DATE JUNE 1977

SUMMARY OF AVERAGE PERCENTAGE SCORES BY OBJECTIVES

TESTS/OBJECTIVES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE	NUMBER OF EXAMINEES
PREREADING	92	87	95	97	87	94	95	97	91	83	86	89	72	70		88.2	244
WORD ATTACK LEVEL 1	95	86	85	77	66	84										82.3	232
WORD ATTACK LEVEL 2	79	69	75	56	81	38	20	12	20	8						45.7	23

PERCENTAGE OF EXAMINEES WHO MASTERED OBJECTIVES

TESTS/OBJECTIVES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE
PREREADING	72	73	90	98	51	80	87	93	81	54	72	84	66	54		76.7
WORD ATTACK LEVEL 1	93	75	78	53	50	81										71.6
WORD ATTACK LEVEL 2	70	52	52	17	57	30	9	0	9	4						30.0

DISTRICT SUMMARY OF TEST RESULTS GRADE 2 DATE JUNE 1977

SUMMARY OF AVERAGE PERCENTAGE SCORES BY OBJECTIVES

TESTS/OBJECTIVES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE	NUMBER OF EXAMINEES
PREREADING	93	82	96	97	86	98	100	100	91	50	88	86	32	69		83.4	11
WORD ATTACK LEVEL 1	97	97	96	94	83	94										93.5	222
WORD ATTACK LEVEL 2	90	86	85	73	83	66	78	57	88	53						75.9	196
WORD ATTACK LEVEL 3	50	63	61	62	67	92	69	19	81	24						58.7	6
WORD ATTACK LEVEL 4	50	98	96	99	81	86	89	88	81	63	0	0	0			64.0	15

PERCENTAGE OF EXAMINEES WHO MASTERED OBJECTIVES

TESTS/OBJECTIVES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE
PREREADING	73	55	91	100	55	91	100	100	73	27	73	82	27	55		71.4
WORD ATTACK LEVEL 1	98	95	95	77	74	92										88.5
WORD ATTACK LEVEL 2	80	69	70	56	71	55	50	16	78	40						58.5
WORD ATTACK LEVEL 3	0	17	33	0	50	83	33	0	50	0						26.7
WORD ATTACK LEVEL 4	0	93	87	93	40	90	80	80	53	40	0	0	0			49.7

OPY AVAILABLE

BEST COPY AVAILABLE

The following breakdown, rather than being district-wide by grade, is by school, across grades within the school. Once again, percentage correct and percentage who mastered each objective on each test is presented. (School names are changed to preserve their anonymity.) The same type of data can be reported for each grade within a school, and also class within a school.

SCHOOL SUMMARY OF PERFORMANCE ON EACH READING OBJECTIVE
FOR EACH TEST ADMINISTERED

Title I Students **School = Green Valley** DATE= JUNE 1977

SUMMARY OF AVERAGE PERCENTAGE SCORES BY OBJECTIVES

TESTS/OBJECTIVES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE	NUMBER OF EXAMINEES
PREREADING	97	97	98	98	88	99	100	100	93	73	78	98	98	93		92.7	73
WORD ATTACK LEVEL 1	88	89	96	78	71	92										85.3	68
WORD ATTACK LEVEL 2	98	91	90	79	89	75	87	75	92	69						86.4	86
WORD ATTACK LEVEL 3	73	65	75	78	73	76	86	57	85	55						72.9	67
WORD ATTACK LEVEL 4	86	83	72	57	83	85	69	81	57	65	90	85	73			75.2	39
DICTIONARY	86	83	77	57	69											78.1	68

PERCENTAGE OF EXAMINEES WHO MASTERED OBJECTIVES

TESTS/OBJECTIVES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE
PREREADING	91	94	94	100	76	94	100	100	87	49	52	87	96	74		85.1
WORD ATTACK LEVEL 1	96	81	91	57	63	81										78.2
WORD ATTACK LEVEL 2	93	93	77	78	83	60	67	46	87	62						72.6
WORD ATTACK LEVEL 3	56	28	51	50	52	58	78	19	69	1						45.7
WORD ATTACK LEVEL 4	62	57	30	16	60	64	36	73	13	37	84	77	53			49.2
DICTIONARY	76	78	73	29	22											55.[illegible]

Titale I Students SCHOOL = **Humbleton** DATE= JUNE 1977

SUMMARY OF AVERAGE PERCENTAGE SCORES BY OBJECTIVES

TESTS/OBJECTIVES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE	NUMBER OF EXAMINEES
PREREADING	100	100	100	100	100	100	92	100	97	100	92	95	100	95		97.9	6
WORD ATTACK LEVEL 1	98	93	96	81	76	85										87.4	14
WORD ATTACK LEVEL 2	93	89	86	71	79	50	71	57	85	55						74.9	14
WORD ATTACK LEVEL 3	88	72	90	82	84	32	91	49	83	51						72.2	15
WORD ATTACK LEVEL 4	89	97	77	77	73	77	93	79	58	92	94	91	73			81.8	11
DICTIONARY	100	96	96	71	53											83.2	11

PERCENTAGE OF EXAMINEES WHO MASTERED OBJECTIVES

TESTS/OBJECTIVES	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	AVERAGE PERCENTAGE PER OBJECTIVE
PREREADING	100	100	100	100	100	100	75	100	100	100	75	100	100	100		96.4
WORD ATTACK LEVEL 1	100	93	93	71	57	79										82.1
WORD ATTACK LEVEL 2	86	56	57	57	57	29	14	7	71	53						49.3
WORD ATTACK LEVEL 3	87	60	73	53	53	60	86	7	80	7						51.3
WORD ATTACK LEVEL 4	82	91	77	45	27	36	82	73	9	45	100	73	55			56.4
DICTIONARY	100	100	100	27	18											69.1

AVAILABLE

BEST COPY AVAILABLE

The final set of tables is a pretest-posttest analysis of one of the six reading tests. Note that there are two tables for each test; one giving percent performance and the other percent mastery. The cells give pretest results (October), posttest results (May), and a percentage gain score. The data is presented for each grade(s) in which the test was administered. Rick DeFriesse, from the Laboratory of Psychometric and Evaluative Research at the University of Massachusetts, Amherst, developed the computer program to produce the table.

PRETEST-POSTTEST ANALYSIS OF THE READING SKILLS INVENTORY RESULTS (1976-1977)

GREEN VALLEY

*** WORD ATTACK LEVEL 2 ***

OBJECTIVE	GRADE 2 (N = 136) % PERFORMANCE			GRADE 3 (N = 165) % PERFORMANCE			GRADE 4 (N = 634) % PERFORMANCE					** ALL GRADES ** (N = 352) % PERFORMANCE		
	OCT	MAY	GAIN	OCT	MAY	GAIN	OCT	MAY	GAIN			OCT	MAY	GAIN
PH3	55	93	38	81	95	14	82	98	16			71	95	23
PH6	65	89	24	78	93	15	84	86	2			74	91	17
PH7	51	86	34	79	94	15	86	93	8			69	91	21
PH8	30	75	46	62	90	28	68	78	11			50	83	33
PH9	48	84	36	76	92	16	85	89	4			67	89	22
PH11	30	69	39	59	80	21	60	70	10			48	75	27
SA1	46	81	35	73	86	13	73	81	8			63	83	21
SA2	16	60	44	43	75	33	54	77	23			34	70	36
SA3	42	91	49	79	94	14	96	93	0			67	92	26
SA4	5	60	54	31	85	54	38	78	40			22	75	53
AVERAGE	39	79	40	66	88	22	72	84	12			57	84	28

* * COMMENTS * *

495

496

BEST COPY AVAILABLE

PRETEST-POSTTEST ANALYSIS OF THE READING SKILLS INVENTORY RESULTS (1976-1977)

GREEN VALLEY

*** WORD ATTACK LEVEL 2 ***

OBJECTIVE	GRADE 2 (N = 132) % MASTERY OCT	MAY	GAIN	GRADE 3 (N = 164) % MASTERY OCT	MAY	GAIN	GRADE 4 (N = 334) % MASTERY OCT	MAY	GAIN	** ALL GRADES ** (N = 747) % MASTERY OCT	MAY	GAIN
PH3	36	[illegible]	54	63	[illegible]	33	55	94	41	54	[illegible]	[illegible]
PH[illegible]	35	77	47	52	89	39	65	74	18	45	83	38
PH7	15	73	58	62	91	31	76	88	12	47	83	[illegible]
PH8	5	62	58	31	88	58	44	65	32	22	75	55
PH9	23	77	54	51	91	43	74	82	12	43	85	[illegible]
PH11	13	59	46	33	76	46	38	47	21	27	66	[illegible]
SA1	5	56	51	34	73	40	32	59	32	24	64	[illegible]
SA2	1	22	21	6	55	52	15	58	35	7	43	[illegible]
SA3	21	86	66	57	89	34	82	79	9	48	95	[illegible]
SA4	2	48	47	12	86	73	21	62	44	9	57	[illegible]
AVERAGE	16	65	51	40	82	44	51	73	26	33	74	44

COMMENTS

BEST COPY AVAILABLE

In sum, there are a wide variety of ways in which the data just presented could be of use to decision makers. We have presented the tables to the reader as an example of a viable method for reporting group test score data.

The next example, taken from Hambleton, Gorth, and O'Reilly (1973), demonstrates how a summary of group performance by objective across test administrations is helpful for decision-making purposes. We present the relevant figure first. The discussion that follows the figure is taken directly from Hambleton, Gorth, and O'Reilly.

Figure 1. Achievement Profiles of A Group of Students on Four Objectives Across Eight Test Administrations.

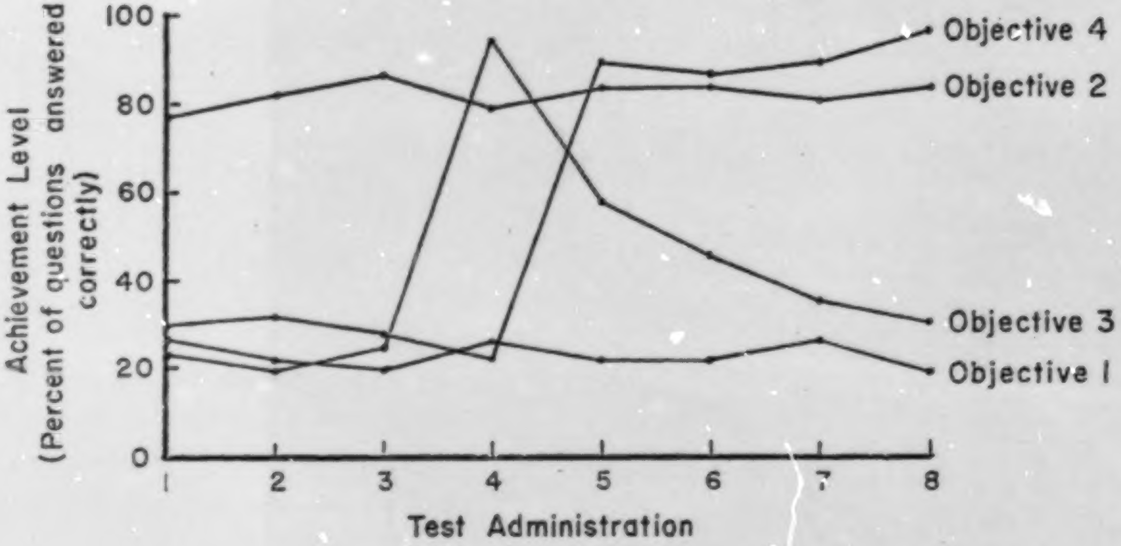


Figure 1 presents hypothetical levels of achievement for four objectives across eight test occasions. In this example , Objective 1 was taught between the first and second test occasion, Objective 3 between the third and fourth test occasion and Objective 4 between the fourth and fifth. For the reason given below, Objective 2 was not taught. On the pretest in the example, all objectives except number 2 show achievement at the chance level or at about 20% on the five-option, multiple-choice items. From an analysis of the data after the second test occasion, the following decisions might be made: (a) Objective 1 was not learned and should probably be retaught in a somewhat different way; (b) since the performance level on Objective 2 was high on both the first and second test occasion, one could safely skip instruction on it. After the sixth test occasion, the following decision could be made on the basis of the data: (a) the performance level on Objective 3 is slipping; if it is an important objective it should be reviewed. It is also noted that the performance level on Objective 1 has not changed. One might postulate that Objective 1 is just too difficult for this particular group of students.

Finally, the table that follows, taken from Millman (1972), may be helpful to the reader when he/she has to decide aobut the number of students needed in a testing situation. The table is self-explanatory.

Table 1[1]

Maximum Percent of Time That A Given Error Will Occur For Selected Test Group Sizes

Number of Students Needed for Testing	Error That Can be Tolerated in Estimating the True Proportion of All Students Who Can Pass An Item				
	10%	15%	20%	25%	30%
10	75[a]	34	34	11	11
15	61	30	12	4	4
20	50	26	12	4	1
25	42	11	4	1	<1
30	36	10	4	1	<1
40	27	8	1	<1	<1
50	20	3	1	<1	<1
60	16	3	<1	<1	<1
75	11	1	<1	<1	<1
100	6	<1	<1	<1	<1
150	2	<1	<1	<1	<1
200	1	<1	<1	<1	<1
250	<1	<1	<1	<1	<1

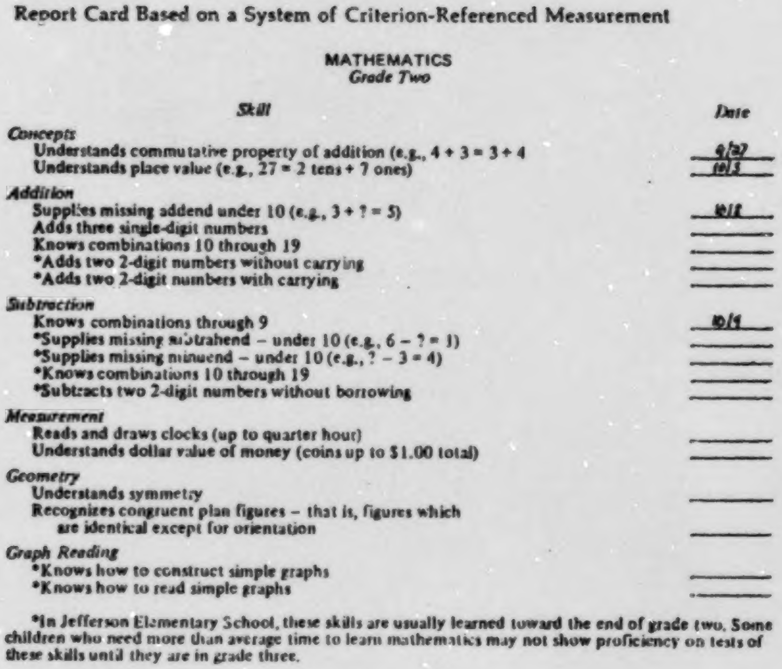
[1] This table is reproduced (with permission and with minor changes) from Millman (1972).

[a] The number "75" has the following interpretation: When a random sample of 10 examinees is used to estimate the proportion of examinees in the population who can answer the item correctly, the likelihood that the estimate will be off at least 10% is no more than .75.

8.7 Grading

In this section of Unit 8, we will discuss two aspects of using criterion-referenced test scores in the grading process. First, we will discuss how one might best grade a student on the activities he/she has undertaken in an objectives-based program. Then, we will discuss the issue of how one assigns final grades in such a program. However, before undertaking this discussion, we'd like to direct the reader to two sources that do an excellent job of comparing and contrasting norm and criterion-referenced grading procedures. These are the 1970 article by Millman in Phi Delta Kappan and the 1974 book by Gronlund on Improving Marking and Reporting in Classroom Instruction.

Since grading in an objectives-based program does not compare students, but rather references the student's performances to the objectives, a single checklist is the best form for grading. If the instruction is group-based, a check mark next to the objectives that were mastered is sufficient. However, if the instruction is individualized, the date of mastery can be placed next to the objective to give a better indication of progress. The following example, taken from Millman (1970), is an example of the latter sort of checklist:



Report Card Based on a System of Criterion-Referenced Measurement

MATHEMATICS
Grade Two

Skill	*Date*
Concepts	
Understands commutative property of addition (e.g., 4 + 3 = 3 + 4	9/27
Understands place value (e.g., 27 = 2 tens + 7 ones)	10/3
Addition	
Supplies missing addend under 10 (e.g., 3 + ? = 5)	10/8
Adds three single-digit numbers	
Knows combinations 10 through 19	
*Adds two 2-digit numbers without carrying	
*Adds two 2-digit numbers with carrying	
Subtraction	
Knows combinations through 9	10/9
*Supplies missing subtrahend – under 10 (e.g., 6 – ? = 1)	
*Supplies missing minuend – under 10 (e.g., ? – 3 = 4)	
*Knows combinations 10 through 19	
*Subtracts two 2-digit numbers without borrowing	
Measurement	
Reads and draws clocks (up to quarter hour)	
Understands dollar value of money (coins up to $1.00 total)	
Geometry	
Understands symmetry	
Recognizes congruent plan figures – that is, figures which are identical except for orientation	
Graph Reading	
*Knows how to construct simple graphs	
*Knows how to read simple graphs	

*In Jefferson Elementary School, these skills are usually learned toward the end of grade two. Some children who need more than average time to learn mathematics may not show proficiency on tests of these skills until they are in grade three.

(Reproduced with permission, from Millman, 1970.)

BEST COPY AVAILABLE

While criterion-referenced testing is highly appropriate for monitoring student progress through the units of instruction making up a course, the question often asked is, "How should final grades in an objectives-based course be assigned?" The issue of course grading has been hotly debated by administrators, teachers, and students. That the issue is important is clear when it is recognized that grades affect career choices of many students and their attitude toward learning, the amount of learning, and the amount of time spent in study. Unfortuantely though, because of the confusion over the purposes of grading and the inexperience of most instructors in areas of tests and measurements, much of final grading is done rather badly. Within objectives-based courses, the purpose of grading is clear and unequivocal. The purpose of final grading is to indicate the overall level of accomplishment of each student relative to the course objectives.

How should a final examination be prepared? One highly acceptable way has been discussed by Block (1971) within the context of mastery learning programs.

The instructor determines the amount of time required for the final examination (often one to two hours) and then proceeds to select test items from the available pools of items measuring the course objectives, preferably items that were not included in any of the unit tests. The items are selected to be representative of the course objectives. Depending on the number of course objectives, and the time available for testing, some course objectives may not be tested in the final examination. The key concern is to develop a final examination such that the test items can be considered to be a representative sample of the material covered in the course.

Let us assume that your particular school insists that letter grades be assigned to students to reflect their work in the course. This constraint should pose no serious problem to the teacher. The test is designed to provide test scores that can be used to infer a student's level of mastery of the course content. The instructor's task is to define the levels of performance that he/she feels reflect A, B, C, D, and F grade level work. For example, the instructor may decide that the appropriate values are 90%, 80%, 70%, and 60%, respectively. These values can be made known to the students and even discussed with them.

Because of the way the test is constructed (sampling of items to be representative of the course objectives), the setting of performance standards can be done on a test score scale that has some real meaning. Certainly, the usual test score scales have little meaning since one can seldom think of the items as a sample from any well-defined domain and therefore the only basis for test score interpretation is to compare one score with another. Grades in objectives-based courses are assigned to students on the basis of their test performance relative to the performance standards that are set to reflect different levels of mastery of course objectives.

The matter of combining unit test score results with final examination results to produce a final grade will not be discussed here, but the problem is a relatively simple one to resolve statistically. Factors such as the relative importance of unit tests versus a final examination would need to be considered in determining the most desirable weighting factors for the two sources of test information.

In sum, it can be seen that the necessity for assigning final grades in a course is amenable to an objectives-based program that utilizes criterion-referenced tests.

8.8 References

Block, J. H. (Ed.) *Mastery learning: Theory and practice*. New York: Holt, Rinehart, and Winston, 1971.

Ebel, R. L. Content standard test scores. *Educational and Psychological Measurement*, 1962, *22*, 11-17.

Ferguson, R. L., & Novick, M. R. Implementation of a Bayesian system for decision analysis in a program of individually prescribed instruction. *ACT Research Report No. 60*. Iowa City, Iowa: American College Testing Program, 1973.

Fremer, J. Handbook for conducting task analysis and developing criterion-referenced tests of language skills. PR 74-12. Princeton, New Jersey: Educational Testing Service, 1974.

Gronlund, N. E. *Improving marking and reporting in classroom instruction*. New York: Macmillan, 1974.

Gronlund, N. E. *Measurement and evaluation in teaching*. (3rd. ed.) New York: Macmillan, 1976.

Hambleton, R. K., & Gifford, J. A. Development and use of criterion-referenced tests to evaluate program effectiveness. *Laboratory of Psychometric and Evaluative Research Report No. 52*. Amherst, MA: School of Education, University of Massachusetts, 1977.

Hambleton, R. K., Gorth, W. P., & O'Reilly, R. P. An application of an evaluation model for classroom instruction. *Journal of Educational Technology Systems*, 1973, *2*, 117-131.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. *Journal of Experimental Education*, 1976, *45*, 57-64.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement*, 1973, *10*, 159-170.

Hambleton, R. K., Swaminathan, H., & Algina, J. Some contributions to the theory and practice of criterion-referenced testing. In D. N. M. de Gruijter, and L. J. Th. van der Kamp (Eds.), *Advances in psychological and educational measurement*. New York: Wiley, 1976.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research*, 1978, *48*, 1-47.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

Huynh, H. Reliability of multiple classifications. Psychometrika, 1978, 43, 317-325.

Jackson, P. H. Simple approximations in the estimation of many parameters. British Journal of Mathematical and Statistical Psychology, 1972, 25, 213-229.

Lewis, C., Wang, M. M., & Novick, M. R. Marginal distributions for the estimation of proportions in m groups. ACT Technical Bulletin No. 13. Iowa City, Iowa: The American College Testing Program, 1973.

Lewis, C., Wang, M. M., & Novick, M. R. Marginal distributions for the estimation of proportions in m groups. Psychometrika, 1975, 40, 63-75.

Linden, W. J., & Mellenbergh, G. J. Optimal cutting scores using a linear loss function. Applied Psychological Measurement, 1977, 1, 593-599.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Livingston, S. A. A utility based approach to the evaluation of pass/fail testing decision procedures. COPA Research Report. Princeton, N.J.: Educational Testing Service, 1975.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Millman, J. Reporting student progress: A case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 226-230.

Millman, J. Determining test length: Passing scores and test lengths for objectives-based tests. Instructional objectives exchange, Los Angeles, California, 1972.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Co., 1974.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.

Popham, W. J. Educational evaluation. Englewood Cliffs, N.J.: Prentice-Hall, 1975.

Swaminathan, H., Hambleton, R. K., & Algina, J. A Bayesian decision-theoretic procedure for use with criterion-referenced tests. Journal of Educational Measurement, 1975, 12, 87-98.

Wang, M. M. Tables of constants for the posterior marginal estimates of proportions in m groups. ACT Technical Bulletin No. 14. Iowa City, Iowa: The American College Testing Program, 1973.

Unit 9

Design of Criterion-Referenced Testing Programs[1]
-Two Examples-

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

---

[1] Substantial portions of the material in this unit were drawn from Hambleton, R.K., Testing and decision-making procedures for selected individualized instructional programs. Review of Educational Research, 1974, 44, 371-400.

Table of Contents

Page

9.0 Overview. 1

9.1 Introduction. 2

9.2 Individualized Instructional Programs 3

9.3 Instructional Models Under Consideration. 5

9.4 Individually Prescribed Instruction (IPI) 6

9.4.1 Instructional Paradigm. 6
9.4.2 Testing Model Description 8
9.4.3 Summary Comments. 14

9.5 Mastery Learning. 15

9.5.1 Instructional Paradigm. 16
9.5.2 Testing Model Description 18
9.5.3 Summary 22

9.6 Summary 23

9.7 References Cited. 24

9.8 References for Further Study. 26

9.0 Overview

Previous units have concentrated on the development, validation, and usage of criterion-referenced tests. In this unit, we will consider two examples where criterion-referenced tests are used to serve a variety of instructional purposes.

9.1 Introduction

The primary purpose of the unit is to introduce readers to the nature of individualized instructional programs and to two testing programs that are in wide use: Individually-Prescribed Instruction (Glaser, 1968) and Mastery Learning (Block, 1971; Bloom, 1976).

9.2 Individualized Instructional Programs

The idea of developing instructional programs in our schools to meet individual student needs is not a new theme in American education (Washburne, 1922), but it has been only since the early 1960's that such programs have been implemented on any large-scale basis in the schools.

The basic argument in favor of individualizing instruction comes from a multitude of research and evaluation studies that suggest that students differ in interests, motivation, learning rate, goals, and capacity for learning, among other things; and, therefore, group-based instruction on a common curriculum is inappropriate to meet their educational needs. The necessity for change in our schools is evident when it is noted, for example, that schools provide successful learning experiences for only one-third of the students (Block, 1971).

The trend toward individualization of instruction in elementary and secondary education and (to a lesser extent) in higher education and technical education, has resulted in the development of a diverse collection of attractive alternative models (Gibbons, 1970; Gronlund, 1974) that, according to their supporters, offer new approaches to student learning that can provide almost all students with rewarding school experiences.

In the relatively short period of time that large-scale individualized instructional programs have been under development, much has been learned about the construction of instructional materials, curriculum design, and computer management (Baker, 1971). However, until recently, corresponding progress was not made in developing relevant testing methods and decision procedures.

One reason for a shortage of testing information was that measurement requirements within the context of many of the new instructional programs required new kinds of tests. These are _criterion-referenced tests_, which are constructed and interpreted in ways quite different from the norm-referenced tests with which most practitioners in the field are familiar. Fortunately, much progress toward a theory and practice of criterion-referenced testing has been made in recent years and many of these developments have been described by Hambleton et al. (1978), Millman (1974), and Popham (1978).

Since one of the major purposes of individualized programs is to maximize the opportunity for all students to learn, it follows that tests used to monitor student progress should be keyed to the instruction presented. Furthermore, they should provide information that can be used to measure progress along an absolute achievement continuum. Norm-referenced tests are constructed specifically to facilitate the making of comparisons among students; hence, they are not very well-suited for making most of the instructional decisions required in individualized instructional programs.

9.3 Instructional Models Under Consideration

Cronbach (1967) discussed three major patterns of dealing with individual differences that provide a framework for the instructional programs considered in this unit. Patterns of dealing with individual differences in schools can be described in terms of the extent to which educational goals and instructional methods are varied. In one pattern, the educational goals and instructional methods are relatively fixed and inflexible. Individual differences are handled mainly by dropping students from a program when they begin to encounter difficulty. In a second pattern, goals are selected for students on the basis of interest and potential, and the students are channeled into one fixed program or another. Individual differences are handled by providing multiple optional programs. Programs described in this unit fit into a third pattern where goals and instructional resources are individualized for the purpose of maximizing learning and development. Although there are hundreds upon hundreds of versions of instructional programs that would fit into this third pattern of individualizing instruction, the two programs we have selected incorporate most, if not all, of the forms of testing that are likely to be found in an individualized instructional program.

Our concern is with individualized instructional programs that include a specification of the curriculum in terms of objectives, detailed diagnosis of the entering competencies of students, the availability of multiple instructional resources, individual pacing and sequencing of material, as well as the careful monitoring of student progress. Thus, our concern is with the most highly structured individualized instructional programs that require substantially more testing than other individual programs, such as the open-classroom plan.

9.4 Individually Prescribed Instruction (IPI)

The Learning Research and Development Center (LRDC) at the University of Pittsburgh initiated the Individually Prescribed Instruction Project during the early 1960's at the Oakleaf School, in cooperation with the Baldwin-Whitehall Public School District near Pittsburgh. As of 1974, the IPI program had been adopted by over 250 schools around the country. We are not aware of any more recent count.

9.4.1 Instructional Paradigm

Although the instructional paradigm and the corresponding test model are discussed in the context of the IPI mathematics program, the procedures, techniques, etc., described, are also applicable for the other content areas covered in the program. In addition, it should be noted that the mathematics program as implemented is probably somewhat different from that described here, since the LRDC is constantly refining and improving the program (Lindvall, personal communication).

Cooley and Glaser (1969) reported that the mathematics curriculum consists of 430 specified instructional objectives. These objectives are grouped into 88 units. (In the 1972 version of the program, there were 359 objectives organized into 71 units.) Each unit is an instructional entity, which the student works through at any one time. There are 5 objectives per unit, on the average, the range being 1 to 14. A collection of units covering different subject areas in mathematics comprises a level; the level may be thought of as roughly comparable to school grades. The number of objectives for each unit in the IPI mathematics curriculum is presented in Table 9.4.1.

Table 9.4.1

Number of Objectives for Each Unit in the IPI Mathematics Curriculum[1]

	Levels							
Content Area	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
Numeration	12	10	8	8	8	3	8	4
Place Value		3	5	10	7	5	2	1
Addition	3	10	5	8	6	2	3	2
Subtraction			4	6	3	1	3	1
Multiplication				8	11	10	6	3
Division				7	7	9	5	5
Combination of Processes			6	5	7	4	5	6
Fractions	3	2	4	6	6	14	5	2
Money		4	4	6	4	1		
Time		3	2	7	9	5	3	1
Systems of Measurement		4	3	5	7	3	2	
Geometry		2	2	3	9	10	7	9
Special Topics			1	3	3	5	4	5

[1] Reproduced by permission from Lindvall, Cox, and Bolvin (1970).

A teacher is faced with the problem of locating, for students, that point in the curriculum where they can most profitably begin instruction. Also, a teacher is responsible for the continuous diagnosis of student mastery as students proceed through their programs of study.

At the beginning of each school year, a teacher places a student within the curriculum; that is, a teacher identifies the units in each content area for which instruction is required. After completing the gross placement, a single unit is selected as the starting point for instruction, and a diagnostic instrument is administered to assess the student's competencies on objectives within the unit. The outcome of the unit test is information appropriate for prescribing instruction on each objective in the unit. In addition, it is also necessary to select the particular set of resources for a student. In theory, resources that match the individual's "learning style" are selected. Within each unit, there are short tests to monitor the student's progress. Finally, upon completion of initial instruction in each unit, assessment and diagnostic testing takes place. In the next section, the tests and the mechanisms for making these decisions are reviewed.

9.4.2 Testing Model Description

Various research reports over the last couple of years have dealt with the testing model and its development (see, for example, Glaser & Nitko, 1971). A flow chart of the testing model is presented in Figure 9.4.1. To monitor a student through the program the following tests are used: Placement tests, unit pretests, unit posttests, and curriculum-embedded tests. All of the tests are criterion-referenced, with performance on the tests compared to performance standards for the purpose of decision-making.

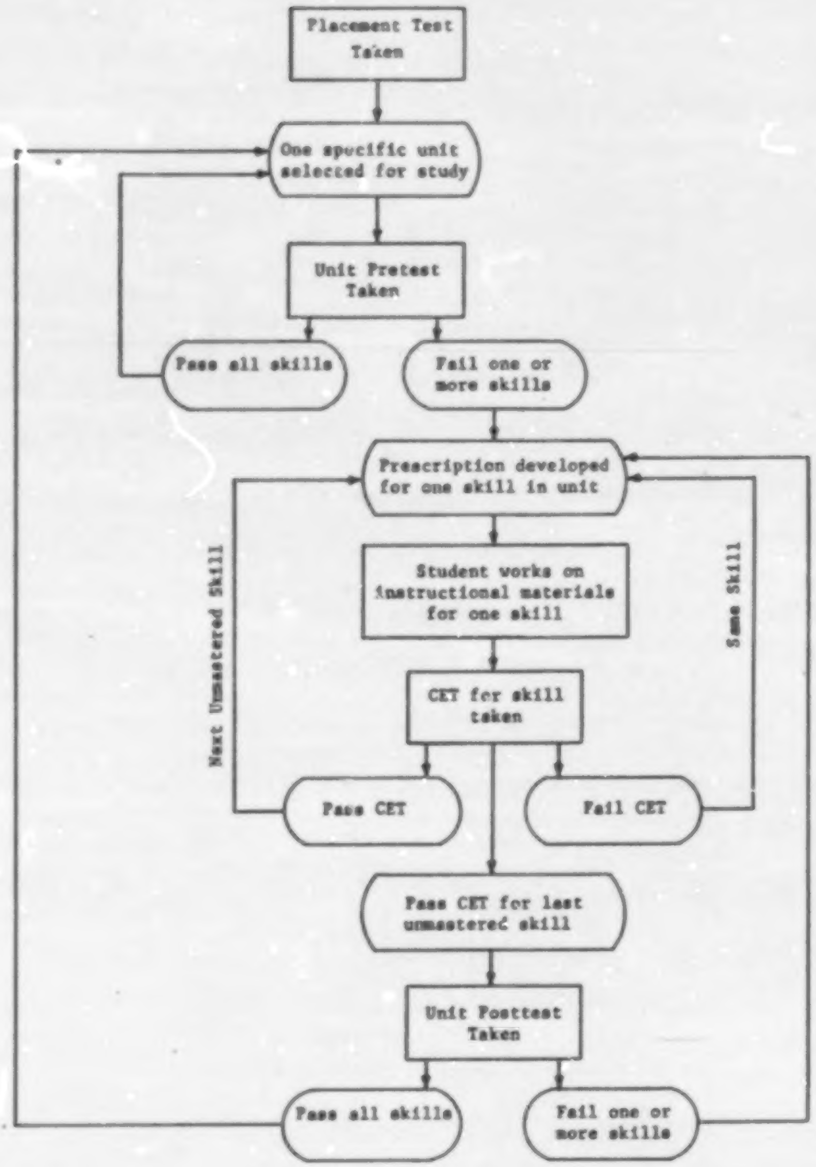

Placement Test Taken
One specific unit selected for study
Unit Pretest Taken
Pass all skills
Fail one or more skills
Prescription developed for one skill in unit
Next Unmastered Skill
Same Skill
Student works on instructional materials for one skill
CET for skill taken
Pass CET
Fail CET
Pass CET for last unmastered skill
Unit Posttest Taken
Pass all skills
Fail one or more skills

Figure 9.4.1. Flow chart of steps in monitoring student progress in the IPI program. (Reproduced, by permission, from Lindvall and Cox, 1969.)

Let us now consider in detail the four kinds of tests and the method for student diagnosis.

Placement Tests. When a new student enters the program, it is necessary to place the student at the appropriate level of instruction in each of the content areas. (Glaser & Nitko, 1971, called this stage-one placement testing.) Typically, this is done by administering a placement test that covers all of the subject areas at a particular level (See Table 9.4.1). Factors affecting the selection of a level for placement testing of a student include student age, past performance, and teacher judgment. Generally, the placement test covers the most difficult or most characteristic objectives within each area. Placement tests are administered until a unit profile identifying a student's competencies within each area is complete. At present, the somewhat arbitrary 80-85% proficiency level is used for most tests in the IPI system.

Student test scores on items measuring objectives in each unit and area in the placement test are used to develop a program of study. The standard procedure is to *assign a student to instruction on units in which placement test performance on items measuring a few representative objectives in the units is between 20% and 80%*. *If the score is less than 20%* for a given unit, *the unit test in the area at the next lowest level is administered* and the same criterion is applied. In the case *where a student has a score of 80% or over, testing the unit in the area at the next highest level is initiated*.

Next we will consider an example. In Table 9.4.2 are shown the test scores of a typical student. The first tests administered to the student are those measuring objectives in Level E. What instruction will be prescribed? What additional testing should be done?

Table 9.4.2

A Set of Criterion-Referenced Test Scores for a Typical Student

	-Level Test-			
Content Area	C	D	E	F
Numeration			60%	
Place Value			90%	60%
Addition			60%	
Subtraction			60%	
Multiplication			30%	
Division			25%	
Combination of Processes			5%	
Fractions			90%	10%
Money			50%	
Time		0%	10%	
Systems of Measurement	85%	40%	0%	
Geometry			30%	
Special Topics			30%	

Example

On the basis of the rules described above, it is likely that the student:

1. would be prescribed instruction at level E in the areas of numeration, addition, subtraction, multiplication, division, combination of processes, money, geometry, and special topics , and
2. would receive the level F placement tests in place value and fractions.
3. If the student scores 60% and 10% in place values and fractions respectively, the student would be assigned to receive instruction at level F in place value and probably level E in fractions.
4. The student would also be administered the level D placement tests in the areas of time and systems of measurement.
5. If the student's scores were 0% and 40% in the areas of time and systems of measurement, respectively, the student would receive a still lower placement test in the area of time and would be prescribed instruction at level D in systems of measurement.
6. If the student scores 85% on the level C placement test in the area of time, the student would be assigned to level D for instruction.

In order to acquire some information on the average length of the tests, the level E placement tests of the 1972 edition of the IPI program were selected and examined. Analysis revealed that, on the average, there are 12 items measuring the objectives in each area (with a range of from six to 20).

In summary, the placement test has the following characteristics: It provides a gross level of achievement for any student in the curriculum, and it provides information for proper placement of students in the curriculum.

Unit Pretests and Posttests. Having received an initial prescription of units, a student proceeds next to take a pretest for a unit at the lowest level of mastery in his/her profile. (Glaser & Nitko, 1971, call this stage-two placement testing.) A unit pretest includes one or more items to measure each objective in the unit. A review of the unit pretests and posttests in level E revealed that the approximate number of items on a test is 37 (the range is from 21 to 64) and the average number of items measuring each objective is six (the range is from four to seven). Lindvall and Cox (1969) report that the length of a pretest is determined by the number of objectives in the instructional unit and by the number of items used to test each objective. No fixed number of items to measure each objective is used because of the diverse nature of the objectives. For example, they note that ". . . an objective like 'the pupil can solve simple addition problems involving all number combinations' will require more items than would an objective like 'the pupil must select which of three triangles is equilateral' (p. 175)."

A student is prescribed instruction in each objective in the unit for which he/she fails to achieve an 85% mastery level of the pretest.[1] In the case where students demonstrate mastery of each objective, they are moved on to the next unit in their profiles, where they again take a pretest.

The unit posttests are simply alternate forms of the unit pretests and are administered to students as they complete instruction on the unit. A student receives a mastery score for each objective in the unit. He/She is required to repeat instruction on any objective where he/she fails to achieve an 85% mastery score. The student is directed to the next unit in his/her profile if he/she demonstrates mastery on each objective covered in the unit posttest. The next unit prescribed is almost always one at the lowest level of mastery (or grade level). Those who repeat instruction on one or more of the objectives must take the unit posttest again before moving on in their program.

---

[1] A mastery score on each objective for a student is calculated as the percentage of items on the test measuring the objective that the student answers correctly.

In summary, pretests and posttests are available for each unit of instruction. The proper pretest is administered on the basis of a student's curriculum profile, and learning tasks for each objective (or skill, as it is called in the IPI program) within the unit are assigned (or not assigned) on the basis of a student's performance on items measuring the objective.

Curriculum-Embedded Tests. As the students proceed through a unit of instruction, their progress is monitored. This is done by the use of curriculum-embedded tests (CET). As used in the mathematics IPI program, a CET is primarily a measure of performance on one specific objective. There are usually several test items to measure the objective. A review of the CETs in level E of the program revealed that there are, on the average, about three items measuring the primary objective covered in the CET. The range is from two to five items. If a student receives a score of 85%, the student is permitted to move on to the next prescribed objective. Otherwise, the student is sent back for additional work before taking an alternate form of the CET.

A second purpose of the CET is to assess, albeit in a fairly crude way, whether or not the student has mastered the next objective in the specified sequence for studying the objectives covered in the unit. If the second objective included in the CET is not one the student has been assigned to study, the student is moved on to be pretested on the second half of a CET that covers the next objective in the student's program of study. Regardless of which CET a student takes, if a score of 85% or over is achieved on the items tested, instruction on the objective is not required. Essentially, this means that a student must score 100% since there are normally only about two items included in the test to cover the second objective. This additional pretesting of an objective in the CET gives students a chance to demonstrate mastery of new skills not specifically covered in the instruction up to that point and to eliminate that instruction from their programs.

Student Diagnosis. Once the student has been assigned to a unit of instruction and the objectives for which instruction is needed have been identified from the unit pretest data, there still remains the problem of deciding which of several instructional methods is "optimal." That is, of the available instructional methods for a particular instructional unit, in which of them would a student with a known background in the program, and specific goals, interests, and aptitudes, stand the "best" chance of learning the material? Glaser and Nitko (1971) call this a diagnostic decision.

9.4.3 Summary Comments

The Individually Prescribed Instruction program is a highly structured system of individualizing instruction that has become a model for literally hundreds of other developers of individualized programs.

9.5 Mastery Learning

The mastery learning concept was introduced to American schools in the 1920's with the work of Washburne (1922) and others in the format of the Winnetka Plan. The program flourished in the 1920's; however, without the technology to sustain a successful program, interest among developers and implementers steadily diminished (Block, 1971). According to Block (1971), mastery learning was revived in the form of programmed instruction in the late 1950's in an attempt to provide students with instructional materials that would allow them to move at their own pace and receive constant feedback on their level of mastery. But programmed instruction was not effective for all students, and so, in an attempt to handle individual differences better, Bloom (1968) and his students (Airasian, 1971; Block, 1971) improved on the standard programmed instruction model by combining it with a model of school learning developed by Carroll (1963, 1970). Carroll's model of school learning provided the conceptual framework for more effective handling of individual differences within an objective-based curriculum. In brief, Carroll's model states that the level of mastery reached by a student on any instructional task or school objective is a function of the time actually spent learning the material and the amount of time the student needs to master the material. The amount of time a student actually spends learning the material depends on two factors—time allowed, and perseverance. The amount of time needed by the student is dependent on three factors—aptitude, quality of the instructional materials, and the student's ability to understand the instructional materials. Carroll goes on to explain how these five factors interact to effect student success in school learning.

Since Bloom's original paper in 1968 describing mastery learning, a considerable amount of mastery learning research has been conducted, and the results suggest that the mastery learning model can be easily and inexpensively implemented in courses at any level of education and in a wide range of content areas (Block, 1970). In particular, Block (1971) notes that the best results have been obtained when the course requires either minimal prior learning, or previous learning, which all or almost all of the students possess. In addition, various research findings have shown better results in courses when the content is highly structured and sequential in nature. The mastery learning model has been used successfully now with more than 100,000 students in elementary, secondary, and college-level courses. The 100,000 figure is a conservative one. Mastery learning programs are being introduced all over the world, and it is no longer possible to keep up with the scope and size of each.

The outstanding features of mastery learning appear to be that it is easily implementable, does not require the use of a computer to manage instruction, and is appropriate for almost any content area. Also, if mastery learning is carried out properly, previous research suggests that students will achieve higher scores and have more interest and a better attitude toward school.

9.5.1 Instructional Paradigm

The curriculum is organized into units of instruction defined by homogeneous clusters of objectives. Initial instruction on the objectives covered in the unit is group-based. In this respect, mastery learning is structurally different from IPI. For each unit, one or more criterion-referenced tests, called formative tests, are used to assess mastery of the objectives. These tests are administered immediately following the completion of the group-based instruction. Individualization is handled via supplemental materials, feedback, and corrective techniques, applied to students who fail to achieve the defined level of mastery on the test items covering the unit objectives. Following the last unit of instruction in the course, a final test covering a representative sample of course objectives is administered, and the data used for grading purposes.

In describing the mastery learning model, Mayo (1970) notes that:

1. Students are made aware of course and unit expectations, so that they view learning as a cooperative rather than as a competitive venture.
2. Standards of mastery are set in advance for the students, and grading is in terms of absolute performance rather than relative performance.
3. Short diagnostic tests are used at the end of each instructional unit.
4. Additional learning is prescribed for those who do not demonstrate unit mastery.
5. Additional time for learning is prescribed to students who seem to need it.

In summary, there are many variations on the basic mastery model, as originally proposed by Bloom (1968). For example, different implementers tend to vary in the extent to which feedback/correction procedures are available and used (Block, 1971). In the next section, the decision points in the program will be considered.

9.5.2 Testing Model Description

Block (1971) notes that "To individualize instruction within the context of ordinary group-based instruction, mastery learning relies heavily on the constant flow of feedback information to teacher and learner (p. 9)." However, it would seem that there is substantially less testing in a mastery learning program than in IPI. A flow chart of the testing model is shown in Figure 9.5.1.

As compared to IPI, there is no placement testing, and unit pretesting and curriculum-embedded testing are not emphasized. Unit posttesting and final assessment represent the two major kinds of testing in the program. Tests to achieve these two purposes are called "formative" and "summative" tests, respectively. Formative tests, or unit posttests as they are called in IPI, are not used for grading. The student data derived from a formative test is used exclusively for diagnosing learning difficulties.

Formative Tests. A formative test, or alternately called a diagnostic-progress test, is a criterion-referenced test that is designed to cover the objectives over a unit of instruction in the mastery learning program. It is used to determine whether or not a student has mastered the material and to serve as a basis for prescribing supplemental work in areas where the student is weak (Airasian, 1971). It is expected also, that the test will reinforce the learning of high-achieving students. Implementers of the mastery learning model have set the passing standard anywhere from 75% to 100%. There is no set number of items or format suggested to measure each objective; in addition, there is a suggestion that instructional decisions are made on the basis of responses to individual items.

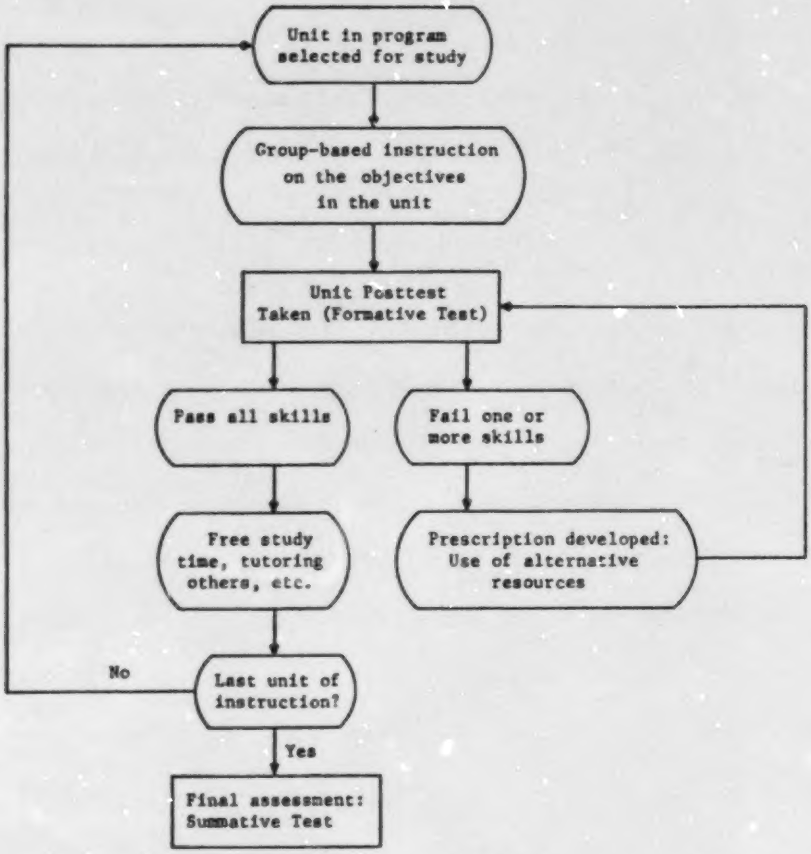

Unit in program selected for study
Group-based instruction on the objectives in the unit
Unit Posttest Taken (Formative Test)
Pass all skills
Fail one or more skills
Free study time, tutoring others, etc.
Prescription developed: Use of alternative resources
No
Last unit of instruction?
Yes
Final assessment: Summative Test

Figure 9.5.1. Flowchart of steps in monitoring student progress in a typical version of a mastery learning program.

The formative tests in mastery learning represent the key to individualizing instruction since it is on the basis of the scores on these tests that individualization of instruction can take place. Units are kept small so that unit testing takes place frequently in order to increase the effectiveness of the individualization of instruction component of the program.

Although it remains an unresolved problem, the matter of setting mastery levels or cutting scores, by which students can be separated into mastery and non-mastery states on the basis of their performance on test items designed to measure objectives included in the criterion-referenced tests, has been more actively researched in the context of the mastery learning program than anywhere else. In addition to the usual concern for setting mastery levels high enough to guarantee that students will have the necessary preparation to begin the next segment of instruction, Block (1970) has noted that, in mastery learning, the mastery level is set in a way that will maximize interest in and attitude toward learning. Some interesting controlled research studies have revealed that a mastery level of about 80-85% is substantially better than a level that is higher or lower. Block's results suggest that setting mastery levels high (95%) may be best for cognitive learning but, in the long run, positive attitudes and interest in the subject are less likely to develop. With a reduction in the mastery level to 85%, there was a reduction in cognitive learning, but selected affective outcomes were maximized. If the mastery level is set lower than 80-85%, students do not usually have sufficient mastery of the skills to proceed effectively with the instruction.

Summative Tests. The primary purpose of the summative test in the mastery learning model is to grade students on the basis of their achievement of course objectives. The items in the test are keyed to objectives and are selected to be representative of the total pool of course objectives. A criterion-referenced interpretation of the scores is recommended. Bloom (1971) proposed that cutting points be located on the ability continuum and that grades should be assigned on the basis of a student's position on the continuum and not relative to other students in the course. A norm-referenced interpretation of the scores is also possible.

Assignment to Instructional Modes. A key part of the mastery learning program is the availability of an extensive number of instructional methods for use by students who fail to demonstrate mastery of the objectives covered on the formative test. A formative test is administered at the end of the group-based instruction on the unit objectives.

Among the alternative resources that are typically available to the student are: Small-group problem sessions, individual tutoring, and alternative learning materials, such as alternate textbooks, workbooks, programmed instruction, audiovisual methods, academic games and puzzles, and reteaching.

The developers of the program have left the decision on the appropriate instructional correctives to the student. It is expected that, through experimentation with many of the instructional correctives, the student will eventually learn which is "best." This would seem to be a very realistic solution to the problem because of the shortage of available data on the appropriate matches between student characteristics and instructional correctives.

9.5.3 Summary

Mastery learning is less different from conventional instruction than IPI since initial instruction on objectives in a mastery learning program is group-based and final grades are assigned. On this latter point, however, it should be noted that because of the organization of the curriculum and the approach to test development and test score interpretation, it is unlikely that the final assessment is as threatening a situation to the student as it usually is in more conventional programs. As compared to conventional instructional programs, mastery learning programs include features such as individual pacing, the frequent use of criterion-referenced tests on small units of instruction to diagnose learning problems, and feedback/corrective techniques.

9.6 Summary

The successful implementation of an individualized instructional program depends, in part, upon the availability of appropriate testing and decision-making procedures to monitor student progress. In this unit we have described and compared the testing models of two of the best known and widely adopted instructional programs: IPI, and Mastery Learning.

9.7 References Cited

Airasian, P. W. The role of evaluation in mastery learning. In J. H. Block (Ed.), *Mastery learning: Theory and practice*. New York: Holt, Rinehart and Winston, 1971.

Baker, F. B. Computer-based instructional management systems: A first look. *Review of Educational Research*, 1971, *41*, 51-70.

Block, J. H. The effects of various levels of performance on selected cognitive, affective, and time variables. Unpublished doctoral dissertation, University of Chicago, 1970.

Block, J. H. (Ed.) *Mastery learning: Theory and practice*. New York: Holt, Rinehart and Winston, 1971.

Bloom, B. S. Learning for mastery. *Evaluation Comment*, 1968, 1(2).

Bloom, B S. Mastery learning. In J. H. Block (Ed.), *Mastery learning: Theory and practice*. New York: Holt, Rinehart and Winston, 1971.

Bloom, B. S. *Human characteristics and instruction: A theory of school learning*. New York: McGraw-Hill, 1976.

Carroll, J. B. A model of school learning. *Teachers College Record*, 1963, *64*, 723-733.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. *Educational Horizons*, 1970, *48*, 71-80.

Cooley, W. W., & Glaser, R. The computer and individualized instruction. *Science*, 1969, *166*, 574-582.

Cronbach, L. J. How can instruction be adapted to individual differences? In R. M. Gagné (Ed.), *Learning and individual differences*. Columbus, Ohio: Charles E. Merrill, 1967.

Gibbons, M. What is individualized instruction? *Interchange*, 1970, *1*, 28-52.

Glaser, R. Adapting the elementary school curriculum to individual performance. In *Proceedings of the 1967 Invitational Conference on Testing Problems*. Princeton, N.J.: Educational Testing Service, 1968.

Glaser, R., & Nitko, A. J. Measurement in learning and instruction. In R. L. Thorndike (Ed.), *Educational measurement*. (2nd ed.) Washington: American Council on Education, 1971.

Gronlund, N. E. *Individualizing classroom instruction*. New York: Macmillan Publishing Co., 1974.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Lindvall, C. M., & Cox, R. The role of evaluation in programs for individualized instruction. In R. W. Tyler (Ed.), Educational evaluation: New roles, new means. Sixty-eight Yearbook, Part II. Chicago: National Society for the Study of Education, 1969.

Lindvall, C. M., Cox, R. C., & Bolvin, J. O. Evaluation as a tool in curriculum development: The IPI evaluation program. AERA Monograph Series on Curriculum Evaluation, No. 5. Chicago: Rand McNally, 1970.

Mayo, S. T. Mastery learning and mastery testing. NCME Measurement in Education, 1970, 1, 3.

Millman, J. Criterion-referenced measurement. In W. J. Popham, (Ed.), Evaluation in education: Current practices. Berkeley, Calif.: McCutchan Publishers, 1974.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Washburne, C. W. Educational measurements as a key to individualizing instruction and promotions. Journal of Educational Research, 1922, 5, 195-206.

9.8 References for Further Study

Block, J. H. (Ed.) Schools, society and mastery learning. New York: Holt, Rinehart, & Winston, 1974.

Block, J. H., & Anderson, L. W. Mastery learning in classroom instruction. New York: Macmillan, 1975.

Bloom, B. S. Human characteristics and instruction: A theory of school learning. New York: McGraw-Hill, 1976.

Davies, I. K. Competency based learning: Technology, management, and design. New York: McGraw-Hill, 1973. (A practical book for teachers providing an introduction to the field of instructional systems.)

Glaser, R. Adaptive instruction: Individual diversity and learning. New York: Holt, Rinehart, and Winston, 1976.

Klausmeier, H. J., Rossmiller, R. A., & Saily, M. Individually guided elementary education: Concepts and practices. New York: Academic Press, 1977. (The book provides an excellent coverage of the theory and practice on Individually Guided Instruction which was developed by the University of Wisconsin Research and Development Center for Cognitive Learning.)

Torshen, K. P. The mastery approach to competency-based education. New York: Academic Press, 1977. (The book provides readers with a good up-to-date review of the theory and research related to competency-based instruction.)

Unit 10

New Developments and Areas for Further Research[1]

Prepared By

Ronald K. Hambleton
University of Massachusetts, Amherst

and

Daniel R. Eignor
Educational Testing Service

March 15, 1979

---

[1] Substantial portions of material in the unit are from Hambleton, R. K., Swaminathan, H., Algina, J., and Coulson, D. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Table of Contents

Page
10.0 Overview of the Unit. 1

10.1 Important Developments and Areas for Further Research and Development. 2

10.2 References. 8

10.0 Overview of the Unit

The purpose of this unit is to introduce practitioners to several important new developments, and to several important criterion-referenced testing topics that have not been satisfactorily resolved.

10.1 Important Developments and Areas for Further Research and Development

One of the most pressing problems for measurement specialists in the 1970's has been the necessity to produce criterion-referenced test technology and instruments—quickly! Unfortunately, the desire of many individuals, organizations, and agencies to use criterion-referenced tests has far exceeded the testing profession's ability to produce test development standards and high quality instruments to meet this need. As a consequence, classroom teachers have been using "home-made" or commercially prepared criterion-referenced tests (which, in most instances, should be called "objectives-referenced tests") of undetermined quality to make instructional decisions; program evaluators (recognizing shortcomings of norm-referenced tests in program evaluation activities) have been constructing criterion-referenced tests based on the "best" principles they can find in a body of literature that is confusing, contradictory, and massive in size (with more unpublished than published papers being circulated); and professional licensing organizations have been grappling with issues such as test score validity and determination of cut-off scores, in the midst of complicated legal actions by the courts. All of the above, as well as many other factors, have contributed to a highly unsettled and volatile situation.

It appears now that there is sufficient theory and practical guidelines for implementing at least adequate criterion-referenced testing programs in situations as far ranging as objectives-based instructional programs at the classroom level, program evaluations at the district and statewide level, and competency-based certification programs at the state and national level.

What important criterion-referenced testing developments and areas for research have emerged? There appear to be several. One, behavioral objectives are being replaced by "amplified objectives" (Millman, 1974) or domain specifications (Popham, 1978). This shift is one of the most important developments because it has implications for the quality of the descriptions that can be made from criterion-referenced test scores. Objectives-referenced tests are being produced by many schools and commercial test publishers, and these tests have value, but they do not permit generalizations from the test scores. Since it is likely that objectives-referenced tests will continue to be produced, it is important for consumers to be familiar with both criterion-referenced tests and objectives-referenced tests and the proper interpretations of scores derived from each type of test. At this stage, there are only a few good examples of domain specifications. These are available from James Popham, Eva Baker and staff at the Center for the Study of Evaluation at UCLA, and Richard Anderson and several of his colleagues at the University of Illinois. Many more domain specifications are under development at various sites around the country, and more will come because the Basic Skills Group at the National Institute of Education has specified the area as one of its priorities.

Two, the role and process of item analysis in test development work seem substantially more clear now. Our review of emerging trends in this area suggests that two types of information should be collected: Item ratings (obtained from any one of many possible formats) of content specialists, and item statistics (of a wide variety of kinds) derived from samples of examinee test item responses. Content specialists need to address two basic questions. One, are the domain specifications clear to potential users and item writers? Two, is the sample of items selected for inclusion in a criterion-referenced test representative of the items defined by a domain specification? On the other hand, item statistics derived from examinee response data may be used to detect "flaws" (for example, technical flaws in items, such as ambiguous wording). A key point emerging from recent literature is that item statistics should not usually be used in item selection since such a strategy introduces a "bias" that could reduce the validity of scores derived from such a test. The one important exception to the rule occurs when the single purpose of a test is to produce scores to make mastery/non-mastery decisions. A better test can be obtained if test items which discriminate in the region of the desired cut-off score are selected.

Three, a significant development is the recognition of the need for construct validation studies with criterion-referenced tests (Linn, 1977; Messick, 1974). The size of the test development project will influence the scope and number of construct validation studies, but clearly more work is needed in this area than has been done in the past. Experimental studies, factor analyses, and investigations of potential sources of low test score validity represent directions for this future research. The limit of these studies will be the level of creativity and ingenuity of the researchers involved (Hambleton, 1977b).

Four, with respect to the technical topics of test length and reliability, there are numerous useful contributions available. More work seems to be needed though with regard to assumptions underlying these technical developments, but generally the work in these areas is sound.

The matter of determining cut-off scores seems less clear at this time (see, for example, Glass, 1978). Aside from the concern about whether cut-off scores should ever be used, at present there are few procedures for sorting through the numerous approaches for determining cut-off scores for the purpose of selecting one. Implementation strategies for nearly all of the approaches are also lacking.

Five, there are numerous Bayesian statistical method contributions offered for improving the precision of domain score estimation and allocating examinees to mastery states. The decision-theoretic procedure outlined earlier provides a framework within which Bayesian statistical methods can be employed with criterion-referenced tests. The incorporation of losses introduces the decision-maker's values into the decision process. The Bayesian methods incorporate the prior knowledge of the decision maker and utilize the data from all examinees, thereby effectively increasing the amount of information the decision maker has without requiring the administration of additional test items. There are a growing number of impressive results to support continued activity in this area (for example, Hambleton, Hutten, and Swaminathan, 1976; Novick and Jackson, 1974; Novick and Lewis, 1974). However, questions about the overall gains that might accrue in view of the complexity of the procedures, the robustness of the Bayesian models in testing situations where the underlying assumptions of the model are not met (for example, when one has very short tests), and the sensitivity of the Bayesian models to the specification of priors, need to be addressed.

Six, a problem which has not been studied at all in the context of criterion-referenced testing, is an instance of the bandwidth-fidelity dilemma (Cronbach and Gleser, 1965). When faced with making a number of decisions of varying importance, and with a limited amount of testing time available, how does a test developer go about determining the "best" distribution of testing time? Does one try to collect considerable test data to make the few most important decisions, or does one try to distribute the available testing time in such a way as to collect a little information relative to each decision? A solution to this problem is required for an efficient testing program. Determination of test lengths for each domain without regard for the size and scope of the total testing program could produce a serious imbalance between testing and instructional time.

Seven, when a set of objectives can be arranged into a learning hierarchy, the strategy of branched testing would seem to offer considerable potential for decreasing the amount of testing while improving its quality (Ferguson, 1969; Hambleton and Eignor, 1977; Spineti and Hambleton, 1977; and Wood, 1973). Some of the practical problems have been resolved in the Pittsburgh IPI Program so that the technique can now be used on a limited basis. Nevertheless, many problems remain before adoption should or can proceed on a large-scale basis. For example, it will be necessary to develop a nonautomated modified version of branched testing for schools without computers. Also, we need to know more about setting starting places, step sizes, stopping rules, etc., before branched testing can be used effectively.

Other matters requiring attention (offered without elaboration) are techniques for reporting criterion-referenced test score information (Ferguson and Novick, 1973; Millman, 1970); the use of norms with criterion-referenced tests (Popham, 1976); applications of latent trait models for the construction of criterion-referenced tests, and evaluations and interpretations of these criterion-referenced test scores; and the nature and scope of training programs for criterion-referenced test developers and users (Hambleton, 1977a).

Consideration was given in our units to topics such as preparing objectives, developing and validating tests, determining reliability, setting cutting scores, and using criterion-referenced test scores. Hopefully, our materials will facilitate the continued development and improvement of criterion-referenced testing. While our list of suggested research and development activities above is not intended to be comprehensive, problem areas suggested above are among the more important ones requiring resolution in the coming years. Our list should be useful as a guide for directing some future work.

In conclusion, there are few criterion-referenced tests available that can meet today's standards for test development, validation, and usage. The good news is that the technology is now sufficiently well-developed to improve this situation. It will be interesting to see what happens.

10.2 References

Cronbach, L. L., & Gleser, G. C. Psychological tests and personnel decisions. (2nd ed.) Washington: American Council on Education, 1971.

Ferguson, R. L. The development, implementation and evaluation of a computer-assisted branched test for a program of individually prescribed instruction. Unpublished doctoral dissertation, University of Pittsburgh, 1969.

Ferguson, R. L., & Novick, M. R. Implementation of a Bayesian system for decision analysis in a program of individually prescribed instruction. ACT Research Report No. 60. Iowa City, Iowa: American College Testing Program, 1973.

Glass, G. V. Criteria and standards. Journal of Educational Measurement, 1978, 15, 237-261.

Hambleton, R. K. What classroom teachers need to know about criterion-referenced testing. Laboratory of Psychometric and Evaluative Research Report No. 50. Amherst, Mass.: School of Education, University of Massachusetts, 1977. (a)

Hambleton, R. K. Validation of criterion-referenced test score interpretations. A paper presented at the Third International Symposium on Educational Testing, University of Leyden, The Netherlands, 1977. (b)

Hambleton, R. K., & Eignor, D. R. Adaptive testing applied to hierarchically structured objectives-based curricula. Proceedings of the Second Computerized Adaptive Testing Conference, University of Minnesota, 1977.

Hambleton, R. K., Hutten, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.

Linn, R. L. Issues of validity in measurement for competency-based programs. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1977.

Messick, S. A. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Millman, J. Reporting student progress: A case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 226-230.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, California: McCutchan Publishing Co., 1974.

Novick, M. R., & Jackson, P. H. Statistical methods for educational and psychological research. New York: McGraw-Hill, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, and W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE monograph series in evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Popham, W. J. Normative data for criterion-referenced tests? Phi Delta Kappan, 1976, 58, 593-594.

Popham, W. J. Criterion-referenced measurement. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

Spinetti, J. P., & Hambleton, R. K. A computer simulation study of tailored testing strategies for objective-based instructional programs. Educational and Psychological Measurement, 1977, 37, 139-158.

Wood, R. Response-contingent testing. Review of Educational Research, 1973, 43, 529-544.